

Cell Press Selections

# Artificial intelligence

Transforming therapeutic development





# See Your Cells in a Whole New Light

VisionSort brings together fundamental advances in optics, microfluidics, and artificial intelligence (AI) to deliver morphological profiling and label-free cell sorting on top of the capabilities you have come to expect from traditional fluorescence-only cytometers. Lose nothing, gain everything with VisionSort.



## VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT

[THINKCYTE.COM](http://THINKCYTE.COM)

## Foreword

---

Today's world is one of abundant data. The generation of high-dimensional biomedical datasets, involving single-cell profiling, high-content imaging, and beyond, has become the new status quo. Meanwhile, population-scale collation of electronic healthcare records provides the opportunity to mine vast amounts of clinical data. In the face of this bounty of information, we collectively wonder how it is possible to understand everything the data are telling us. Enter artificial intelligence (AI). Therapeutic development crosses scientific boundaries, encompassing basic biology, chemical biology, and clinical care. It is particularly amenable to AI, which has the potential to integrate data from these different realms in order to bring effective therapies to the clinic more efficiently.

We are pleased to present this collection of articles from Cell Press focused on applications of AI in therapeutic development. Several present applications of predictive AI for drug discovery, drug synergy, or drug toxicity. Drug discovery is a potentially powerful application of generative AI, and we are including a review focused on generative molecular design for this purpose. And as the ultimate goal is the clinic, we are also including a review focused on the use of machine learning across the spectrum of cancer care.

We hope these papers spark your enthusiasm for the potential of AI for improving therapeutic development and patient outcomes. Finally, we would like to thank ThinkCyte for providing the support that made publication of this collection possible.

### **Ruth Zearfoss**

Editor-in-Chief, *Cell Reports Methods*

#### **For more information about Cell Press Selections:**

Jim Secretario  
Program Director  
j.secretario@elsevier.com  
917-678-0541  
@CellPressBiz





# Resolve the complexity of disease through morphology

Achieve unparalleled new insight into disease with morphological profiling and AI. Uncover and isolate rare and unique cell populations, identify new drug targets and drug resistance mechanisms, or incorporate morphology into patient stratification models. VisionSort elevates biomarker discovery campaigns to a new level with flexible, easy-to-use, and user-controlled AI algorithms embedded directly in the instrument. Discover more with unbiased morphometric characterization.



## VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT  
[THINKCYTE.COM](http://THINKCYTE.COM)

# Artificial intelligence

Transforming therapeutic development

## Opinion

AI-powered therapeutic target discovery

*Frank W. Pun, Ivan V. Ozerov, and Alex Zhavoronkov*

## Reviews

Deep generative molecular design reshapes drug discovery

*Xiangxiang Zeng, Fei Wang, Yuan Luo, Seung-gu Kang, Jian Tang, Felice C. Lightstone, Evandro F. Fang, Wendy Cornell, Ruth Nussinov, and Feixiong Cheng*

From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment

*Kyle Swanson, Eric Wu, Angela Zhang, Ash A. Alizadeh, and James Zou*

## Articles

A hybrid deep forest-based method for predicting synergistic drug combinations

*Lianlian Wu, Jie Gao, Yixin Zhang, Binsheng Sui, Yuqi Wen, Qingqiang Wu, Kunhong Liu, Song He, and Xiaochen Bo*

Machine learning prediction of side effects for drugs in clinical trials

*Diego Galeano and Alberto Paccanaro*

TriNet: A tri-fusion neural network for the prediction of anticancer and antimicrobial peptides

*Wanyun Zhou, Yufei Liu, Yingxin Li, Siqi Kong, Weilin Wang, Boyun Ding, Jiyun Han, Chaozhou Mou, Xin Gao, and Juntao Liu*

## Short article

A deep learning platform to assess drug proarrhythmia risk

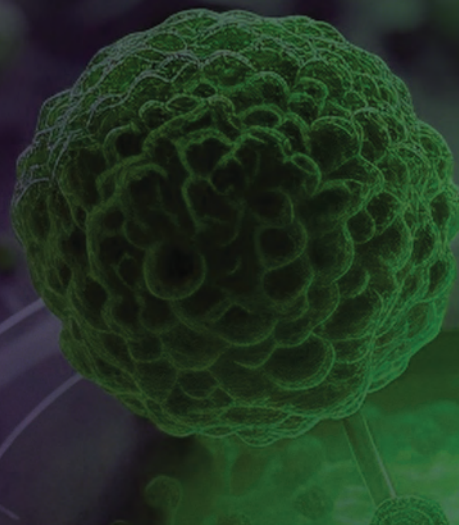
*Ricardo Serrano, Dries A.M. Feyen, Arne A.N. Bruyneel, Anna P. Hnatiuk, Michelle M. Vu, Prashila L. Amatyia, Isaac Perea-Gil, Maricela Prado, Timon Seeger, Joseph C. Wu, Ioannis Karakikes, and Mark Mercola*



## Ditch the Labels.

Optimize cell therapy R&D  
with the power of morphology.

Expand the potential of your cell therapy R&D workflows with label-free cell sorting. Isolate immune cells of interest, select therapeutically meaningful cells, or monitor quality attributes for manufacturing QC...all without using external labels or markers. See how morphological profiling and AI can help optimize the therapeutic potential of your cellular therapies.



### VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT

[THINKCYTE.COM](http://THINKCYTE.COM)

## Report

**RECOVER identifies synergistic drug combinations *in vitro* through sequential model optimization**

*Paul Bertin, Jarrod Rector-Brooks, Deepak Sharma, Thomas Gaudelet, Andrew Anighoro, Torsten Gross, Francisco Martínez-Peña, Eileen L. Tang, M.S. Suraj, Cristian Regep, Jeremy B.R. Hayter, Maksym Korablyov, Nicholas Valiante, Almer van der Sloot, Mike Tyers, Charles E.S. Roberts, Michael M. Bronstein, Luke L. Lairson, Jake P. Taylor-King, and Yoshua Bengio*

## Resources

**Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction**

*Markus Müller, Florian Huber, Marion Arnaud, Anne I. Kraemer, Emma Ricart Altimiras, Justine Michaux, Marie Taillandier-Coindard, Johanna Chiffelle, Baptiste Murgues, Talita Gehret, Aymeric Auger, Brian J. Stevenson, George Coukos, Alexandre Harari, and Michal Bassani-Sternberg*

**Integrating inflammatory biomarker analysis and artificial-intelligence-enabled image-based profiling to identify drug targets for intestinal fibrosis**

*Shan Yu, Alexandr A. Kalinin, Maria D. Paraskevopoulou, Marco Maruggi, Jie Cheng, Jie Tang, Ilknur Icke, Yi Luo, Qun Wei, Dan Scheibe, Joel Hunter, Shantanu Singh, Deborah Nguyen, Anne E. Carpenter, and Shane R. Horman*



**On the cover:** Advances in artificial intelligence are transforming the way therapeutics are developed, making the process faster and cheaper and bringing treatments more efficiently to the clinic. The cover image depicts therapeutic molecules and a capsule for delivery. Image courtesy of MF3d/Getty Images.



# Phenotypic Screening Unleashed with Morphology

Add a new dimension to your phenotypic screens with the exhaustive capacity of morphological profiling. Compatible with both small molecule screens and CRISPR-based approaches, VisionSort opens the door to a wider range of phenotypes to screen. Discover new drugs and drug targets quickly and more efficiently with VisionSort.



## VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT

[THINKCYTE.COM](http://THINKCYTE.COM)



## Opinion

# AI-powered therapeutic target discovery

Frank W. Pun,<sup>1</sup> Ivan V. Ozerov,<sup>1</sup> and Alex Zhavoronkov<sup>1,2,3,\*</sup>

**Disease modeling and target identification are the most crucial initial steps in drug discovery, and influence the probability of success at every step of drug development. Traditional target identification is a time-consuming process that takes years to decades and usually starts in an academic setting. Given its advantages of analyzing large datasets and intricate biological networks, artificial intelligence (AI) is playing a growing role in modern drug target identification. We review recent advances in target discovery, focusing on breakthroughs in AI-driven therapeutic target exploration. We also discuss the importance of striking a balance between novelty and confidence in target selection. An increasing number of AI-identified targets are being validated through experiments and several AI-derived drugs are entering clinical trials; we highlight current limitations and potential pathways for moving forward.**

### Overview of target identification

The drug discovery pipeline is widely recognized to be a time-consuming, expensive, and risk-laden process that typically requires around 10 years and \$2 billion to bring a novel drug to market [1]. By 2022 fewer than 500 successful drug targets had been identified [2], representing a tiny fraction of the estimated druggable targets in humans [3,4]. Although numerous drug candidates undergo extensive optimization during preclinical stages, the average failure rate in clinical trials from 2009 to 2018 reached 84.6%<sup>i</sup>. The lack of clinical efficacy remains the key factor contributing to the failure of both Phase 2 and 3 trials [5], leading to substantial financial losses and resource wastage. Identifying the right drug targets is crucial for increasing the likelihood of developing clinically effective therapies.

Target identification, the process of identifying the right biological molecules or cellular pathways that can be modulated by drugs to achieve therapeutic benefits, is increasingly important in modern drug discovery. Although innovations in experimental and omic technologies have been growing over the past few decades (Figure 1), identifying actionable therapeutic targets remains challenging. The integration of multiomic data with **AI** (see Glossary) algorithms has recently emerged as a promising approach for target identification<sup>ii,iii</sup>. We discuss here the conventional target identification approaches with a focus on the application of AI algorithms to target identification. This paper aims to offer a progressive outlook on the emergence of the AI-driven drug discovery era and encourage the integration of AI technologies into drug discovery pipelines.

### Strategies in target identification: from experiments to machine learning

Target identification can be classified into three distinct strategies – experimental, multiomic, and computational approaches (Figure 2). Using these methods collaboratively can generate novel therapeutic hypotheses in exploratory target identification, thus significantly enhancing our understanding of complex diseases.

### Highlights

Disease modeling and target discovery are crucial initial steps in the drug discovery process and significantly impact on the success of drug development.

Given the advantages of analyzing large datasets and complex biological networks, artificial intelligence (AI) is playing a growing role in modern drug target identification.

We discuss the use of deep learning models for target discovery, AI-identified targets validated through experiments, and the use of synthetic data produced using generative AI for target identification.

Novelty, in addition to druggability and toxicity, is a crucial factor in target selection. There is a trade-off between choosing high-confidence and novel targets.

Over the past few years several AI-derived drugs have entered clinical trials, signaling the dawn of a new era in AI-driven drug discovery.

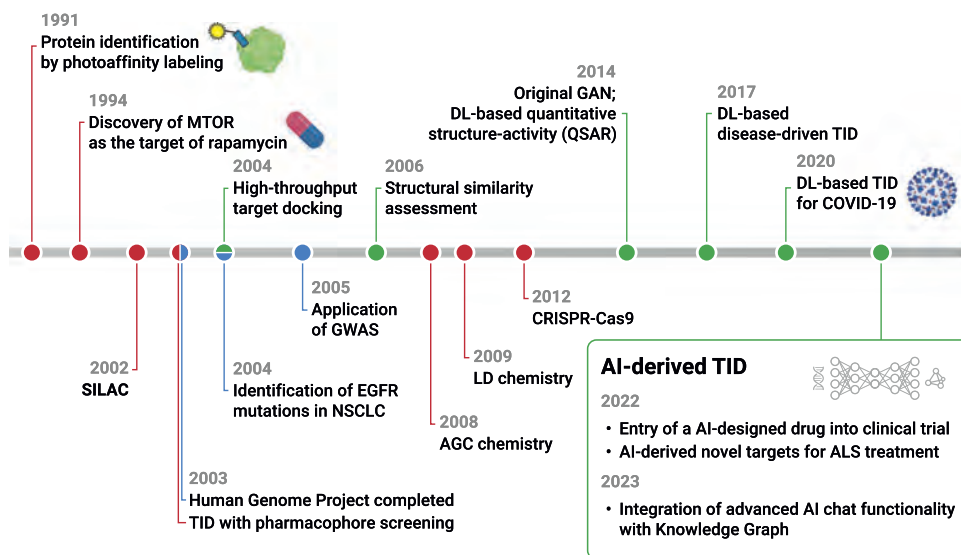
<sup>1</sup>Insilico Medicine Hong Kong Ltd., Hong Kong Science and Technology Park, New Territories, Hong Kong

<sup>2</sup>Insilico Medicine MENA, 6F IRENA Building, Abu Dhabi, United Arab Emirates

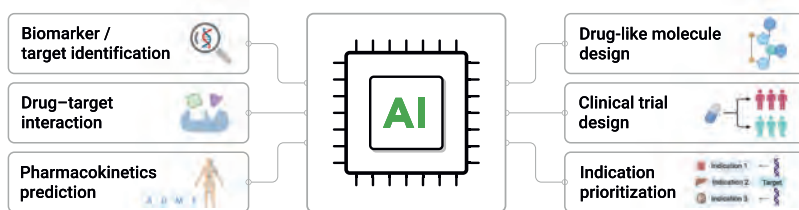
<sup>3</sup>Buck Institute for Research on Aging, Novato, CA, USA

\*Correspondence:  
alex@insilico.com (A. Zhavoronkov).





## AI applications in the early stages of drug discovery



Trends in Pharmacological Sciences

**Figure 1. The emergence of artificial intelligence (AI) in early drug development.** (Upper panel) Key technological advances in the history of target identification are classified into three types: experiment-based (red), multiomic (blue), and computational (green) approaches. Traditionally, experiment-based methods have been the go-to approach for discovering therapeutic targets. However, with the rise of big data, integrated analysis of multiomic data has become a more efficient strategy for target identification. In addition, recent advances in AI-driven biological analysis have identified novel targets and AI-designed drugs are now entering clinical trials. (Lower panel) AI applications in the early stages of drug discovery. Abbreviations: AGC chemistry, affinity-guided catalyst chemistry; ALS, amyotrophic lateral sclerosis; DL, deep learning; EGFR, epidermal growth factor receptor; GAN, generative adversarial network; GWAS, genome-wide association study; LD chemistry, ligand-directed chemistry; MTOR, mammalian target of rapamycin; NSCLC, non-small cell lung cancer; SILAC, stable isotope labeling with amino acids in cell culture; TID, target identification. Figure created with BioRender.com.

### Experimental approaches

Experimental approaches, including affinity-based biochemical, comparative profiling, and chemical/genetic screening, have demonstrated their striking contributions to target identification since the 1960s. The use of small-molecule affinity probes, which allow traceless protein labeling upon ligand–protein interaction [6], is the most straightforward method among the three experimental approaches. The selection of probes is highly dependent on the identity of the starting molecule [7]. Stable isotope labeling by amino acids in cell culture (SILAC), an example of comparative profiling, is a popular quantitative proteomics tool that uses stable isotope-labeled amino acids to accurately differentiate cellular proteomes [8]. Studies conducted in multiple cancer types such as hepatocellular carcinoma (HCC) [9], multiple myeloma [10,11], endometrial cancer [12], and colorectal cancer [13,14] have clearly exemplified the effectiveness of SILAC in identifying pivotal players in disease pathogenesis. Chemical/genetic screening, implemented by RNA interference

### Glossary

**Artificial intelligence (AI):** the ability of a computer or computer-controlled machine to perform problem-solving and decision-making tasks that are commonly associated with intelligent beings.

**Biomarker:** a biological molecule in any type of body fluid or tissue that serves as a sign of a biological state.

**Drug repurposing:** the process of identifying a novel therapeutic application for existing drugs that have been FDA-approved or clinically investigated for specific medical indications.

**Drug–target interaction:** an important step in drug discovery that recognizes how a chemical compound and a protein target interact in the human body.

**Generative adversarial networks (GANs):** a class of machine learning frameworks that consists of two neural networks that compete against each other during the training process and improve their functionalities to generate samples indistinguishable from the real data.

**Genome-wide association study (GWAS):** a method to identify genomic variants that are statistically associated with a risk for a disease or a trait by comparing the frequencies of genomic variants between people with and without that specific disease or trait.

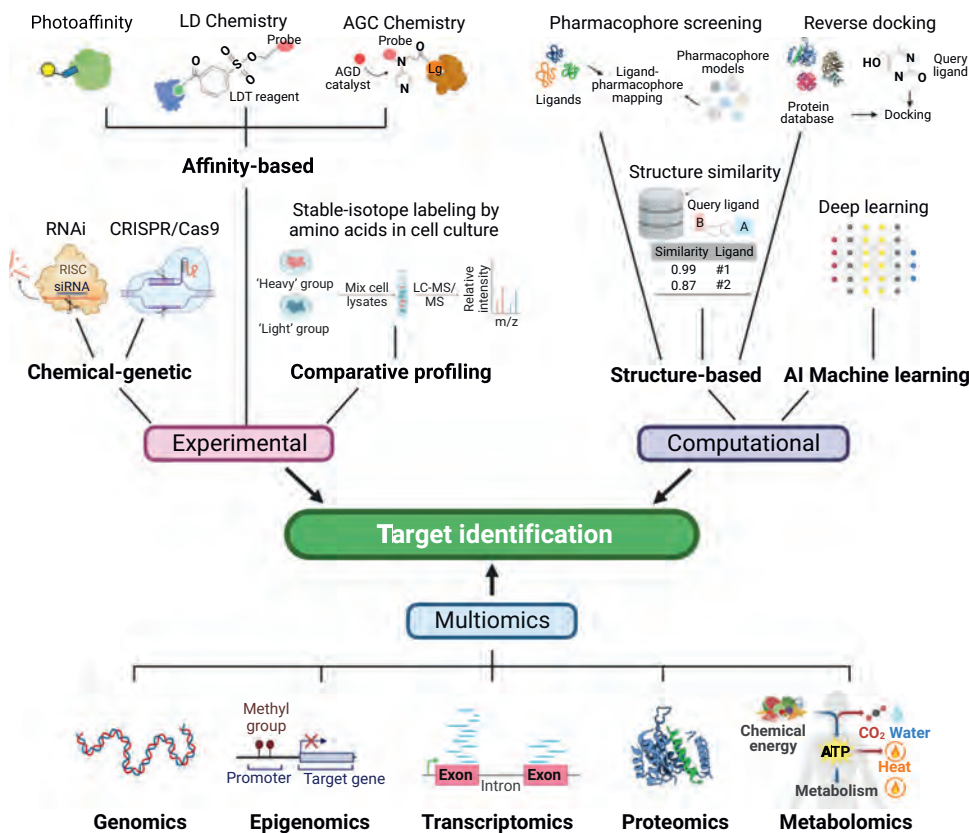
**Indication prioritization:** the process of prioritizing the potential indications of a drug based on the expected relevancy of the drug and a specific indication using AI.

**Induced pluripotent stem cells (iPSCs):** artificial stem cells generated from an adult somatic cell through the coexpression of specific pluripotency-associated genes, namely *c-Myc*, *Oct3/4*, *Sox2*, and *Klf4*.

**Machine learning:** a branch of artificial intelligence that focuses on mimicking human learning processes via the use of data and algorithms to gradually improve its accuracy.

**Natural language processing:** a field of AI that processes and analyzes large amounts of natural language data with a goal to enable computers to understand, interpret, generate human language, and extract information from documents.

**Pharmacokinetics:** the study of the fate of an administered substance in an organism, namely absorption, distribution, metabolism, and excretion.



**Recurrent neural networks:** a class of artificial neural networks with feedback connections that are designed to learn sequential or time-varying data.

**Transfer learning:** a machine learning method where a pretrained model is reused as the starting point for a model on another related task; this approach is commonly used as an optimization technique to save time and increase performance.

**Therapeutic modality:** the type of therapy used to treat a disease or medical condition, including small-molecule drugs, protein-based therapies, advanced therapies (such as cell and gene therapies), and microorganism-based therapies.

**Figure 2. Three exploratory strategies for target identification.** Exploratory techniques for target identification can be classified into three strategies: experimental, multiomic, and computational approaches. The experimental approach involves conducting wet-lab experiments to identify targets based on affinity, genetic modification screening, and comparative profiling. The multiomic approach predicts gene–disease associations by analyzing various omic datasets such as genomics, transcriptomics, proteomics, epigenomics, and metabolomics. Lastly, the computational discovery approach efficiently identifies potential targets by using machine learning or structure-based methods including reverse docking, pharmacophore screening, and structure similarity analysis. Abbreviations: AGC chemistry, affinity-guided catalyst chemistry; AGD, affinity-guided DMAP (4-dimethylaminopyridine); AI, artificial intelligence; LC, liquid chromatography; LD chemistry, ligand-directed chemistry; LDT, ligand-directed tosyl; MS, mass spectrometry; RISC, RNA-induced silencing complex; RNAi, RNA interference; siRNA, short interfering RNA. Figure created with BioRender.com.

(RNAi) or CRISPR-Cas9 gene editing, has been of great interest to biologists for decades. Owing to its high specificity and efficiency [15], CRISPR has dramatically expanded our knowledge of the mechanistic and pharmacological aspects of human diseases. For example, BRD2 was identified as an essential regulator of the host response to SARS-CoV-2 infection by a targeted CRISPR interference screen [16]. Making use of the CRISPR interference- and CRISPR activation-based functional genomics platform, Ramkumar *et al.* identified the determining roles of HDAC7 and the Sec61 complex in modulating the immunotherapy response in multiple myeloma [17]. Although it has been 10 years since its introduction, CRISPR technology continues to evolve to further enhance its flexibility, simplicity, and efficiency, thus offering a great benefit to the research community not only for target identification but also as a gene therapy and diagnostic tool.

### Multiomic approaches

Multiomic data provide researchers with interconnected molecular information from different perspectives, including static genomic data and spatiotemporally dynamic expression and metabolic

profiles [18]. As the first established and most mature omics discipline [19], genomics focuses on genetic variants in the DNA sequence. Large-scale **genome-wide association study (GWAS)** analysis powered by next-generation sequencing has yielded hundreds of thousands of associations between genetic variants and complex diseases or traits [20], leading to the development of breakthrough therapies such as the cystic fibrosis modulator drugs targeting CFTR mutations [21], and novel drugs for the treatment of inflammatory bowel disease targeting the disease-associated gene *IL23A* [22]. More recently, meta-analyses of published GWAS data have revealed novel genetic loci attributable to different diseases, thus opening up **drug repurposing** opportunities [23,24]. Although genomic evidence has been one of the indispensable factors in target identification, distinguishing the causative genetic variants that lead to a given disease remains challenging. In this regard, integrating multiple omic lines of evidence can be useful. Transcriptomic and proteomic data can be used to identify causal genetic loci that regulate gene and protein levels and facilitate the discovery of genes and pathways underlying disease pathogenesis [25–27]. Likewise, epigenomic and metabolomic data can also serve as functional evidence for GWAS-identified variants to support their disease associations and clinical applications [28–30]. As compared to single omic approaches, integrated multiomic analysis can provide a more comprehensive view of disease mechanisms and is therefore increasingly used to facilitate **biomarker** and therapeutic target discoveries, treatment response, and patient prognosis predictions [31–34].

#### Computational approaches

Because typical experiment-based target identification is laborious and resource-intensive, computational approaches have emerged as promising alternatives for achieving efficient target screening. Depending on the availability of protein structure and the chemical structure of the compound of interest, pharmacophore screening [35], reverse docking [36], and structure similarity assessment [37,38] have been used to predict novel biological targets for small molecules. On the other hand, AI is a growing discipline in computational science for target discovery. **Machine learning** is an indispensable component of AI that can be applied either with or without supervision. Supervised learning utilizes labeled datasets to train models for data classification and reliable outcome prediction. By contrast, unsupervised learning explores the hidden structure of unlabeled data without human intervention [39]. The application of machine learning is not limited to predicting biological targets of the existing drugs or compounds, and can also identify novel therapeutic targets for any disease of interest. The details of how machine learning facilitates target discovery for disease treatment will be elaborated upon in the following AI sections.

#### AI-driven target identification

In recent years we have witnessed an explosion of biomedical data ranging from basic research on disease mechanisms to clinical investigation in patients. Although large amounts of information have been generated, the growth of data also poses challenges for data analysis. This is where the emerging role of AI comes into play. Given the advantage of AI in processing and tackling complex biomedical networks of data, using AI algorithms can reveal patterns and relationships within the data that may not be apparent to humans, and may possibly lead to better understanding and treatment of diseases. AI has made notable contributions that facilitate biomarker and target identification [40–42], **indication prioritization** [43], drug-like molecule design [44,45], **pharmacokinetics** prediction [46], **drug–target interaction** [47,48], and clinical trial design [49] (Figure 1, lower panel). Although still in the early stages of clinical trials, AI-derived drugs are increasingly emerging in clinical studies (Table 1), as exemplified by GS-0976 for the treatment of non-alcoholic steatohepatitis, EXS-21546 for solid tumors, and INS018\_055 for idiopathic pulmonary fibrosis, which is the first-ever AI-derived drug with positive topline results in a Phase 1 clinical trial.

Table 1. AI-derived drugs in clinical trials

Company	Target	Indication <sup>a</sup>	Compound	Development status	Trial number <sup>b</sup>
BenevolentAI	Trk	Atopic dermatitis	BEN-2293	Phase 2	NCT04737304
Exscientia	A2AR	Solid tumors	EXS-21546	Phase 1	NCT04727138
	5-HT1A	Obsessive compulsive disorder	DSP-1181	Phase 1	Undisclosed <sup>vi</sup>
	5-HT1A/2A	Alzheimer's disease psychosis	DSP-0038	Phase 1	Undisclosed <sup>vii</sup>
	PKC- $\theta$	Inflammatory diseases	EXS4318	Phase 1/2	Undisclosed <sup>viii</sup>
Insilico Medicine	Target X	Idiopathic pulmonary fibrosis	INS018_055	Phase 2	NCT05938920, CTR20230776
	3CLPro	COVID-19	ISM3312	Phase 1	CTR20230768
	USP1	BRCA-mutant cancer	ISM3091	Phase 1	NCT05932862
Nimbus Therapeutics	ACC	Nonalcoholic steatohepatitis	NDI-010976/GS-0976	Phase 2	NCT02856555, NCT03987074, NCT02891408, NCT02876796
Pharos iBio	FLT3	Acute myeloid leukemia Ovarian cancer Triple-negative breast cancer Radiation sensitizer	PHI-101	Phase 1	NCT04842370 NCT04678102
Recursion Pharmaceuticals	CCM2	Cerebral cavernous malformation	REC-994	Phase 2	NCT05085561
	HDAC	Neurofibromatosis type 2	REC-2282	Phase 2/3	NCT05130866
	MEK1/2	Familial adenomatous polyposis	REC-4881	Phase 2	NCT05552755
Relay Therapeutics	SHP2	Solid tumors	RLY-1971/RG-6433	Phase 1	NCT04252339
	FGFR2	FGFR2-driven cancers Intrahepatic cholangiocarcinoma Advanced solid tumors	RLY-4008	Phase 1/2	NCT04526106
	PI3K $\alpha$	Solid tumors	RLY-2608	Phase 1	NCT05216432
Schrödinger	MALT1	Non-Hodgkin's lymphoma	SGR-1505	Phase 1	NCT05544019
Structure Therapeutics	GLP1R	Type 2 diabetes Obesity	GSBR-1290	Phase 1	NCT05762471
	APLNR	Pulmonary arterial hypertension Idiopathic pulmonary fibrosis	ANPA-0073	Phase 1	ACTRN12621000644864
Valo Health	S1P1	Post-myocardial infarction Acute kidney injury	OPL-0301	Phase 2	NCT05327855
	ROCK1/2	Diabetic retinopathy Diabetic complications	OPL-0401	Phase 2	NCT05393284

<sup>a</sup>Indications retrieved from the company pipeline.

<sup>b</sup>For undisclosed trial numbers, press releases are provided as the source of reference.

### Application of deep learning models in target discovery

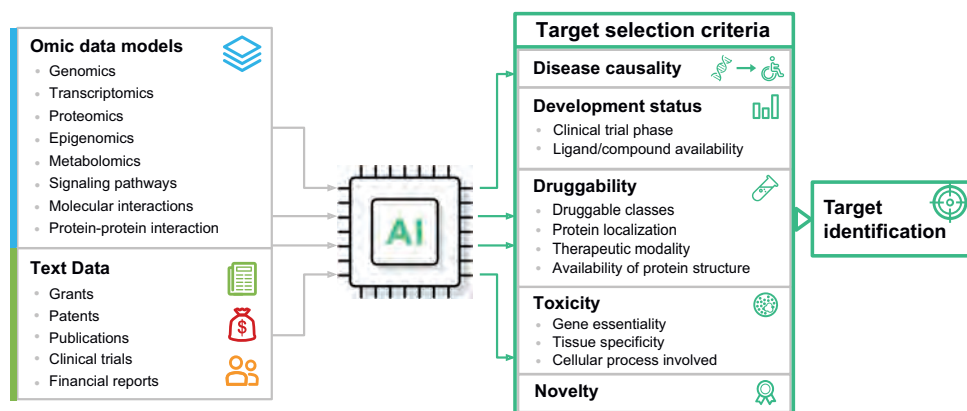
In recent years machine learning-based algorithms, particularly deep learning methodologies, have drawn significant attention and have achieved excellent results in pharmaceutical areas. Deep learning, also known as deep neural networks, consists of multiple hidden layers of nodes through which data processing and feature extraction are conducted successively in a cascade manner [50]. Compared to traditional machine learning methods, more recent deep learning-based architectures, such as **generative adversarial networks (GANs)**, **recurrent**

**neural networks**, and **transfer learning** techniques, have attracted increasing attention and have been applied to various aspects of healthcare, such as *de novo* small-molecule design [51], aging research [44], and pharmacological prediction of drugs based on transcriptional data of drug-perturbed cell lines [52]. Using publicly available multiomic data and text mining (Figure 3, Key figure), deep learning has recently been used in studies of fatal disorders with urgent and unmet clinical needs. To identify actionable therapeutic targets in amyotrophic lateral sclerosis (ALS), Pun *et al.* combined a variety of bioinformatic- and deep learning-based models that were trained using disease-specific multiomic and text-based data to prioritize druggable genes, revealing 18 potential targets for ALS treatment [53]. In addition, Fabris *et al.* established a deep learning-based method with a novel modular architecture to identify human genes associated with multiple age-related diseases by learning patterns retrieved from gene or protein features such as Gene Ontology terms, protein–protein interactions, and biological pathways [54]. West *et al.* developed a deep learning ensemble trained using the transcriptomic profiles of >12 000 embryonic and adult cells [55]. A novel target (COX7A1) for controlling the embryonic–fetal transition was revealed, which could facilitate our understanding of normal development, epimorphic tissue regeneration, and cancer.

Furthermore, large language models also aid therapeutic target discovery via rapid biomedical text mining. Pretrained on a vast amount of text data extracted from millions of publications, large language model-based Chat functionalities, such as BioGPT from Microsoft [56] and ChatPandaGPT from Insilico Medicine<sup>V</sup>, can connect diseases, genes, and biological processes to allow rapid identification of the biological mechanisms involved in disease development and progression, as well as the identification of potential drug targets and biomarkers. The ability of

### Key figure

#### Workflow of artificial intelligence (AI)-driven target discovery



Trends in Pharmacological Sciences

**Figure 3.** AI prioritizes targets for specific indications by using multi-models that utilize a diverse range of publicly available omic and text data. Omic data encompass genomics, transcriptomics, proteomics, epigenomics, and metabolomics. These data provide information about altered signaling pathways, molecular interactions, and protein–protein interactions that can serve as additional inputs for target prioritization. Text-based data are retrieved from funding reports, patents, publications, and clinical trials. During target prioritization, multiple target selection criteria such as protein family class, development status, druggability, toxicity, and novelty can be applied to refine the list of AI-driven targets to align with specific research objectives.

the large language models to understand natural language and interpret complex scientific concepts could make them valuable tools in accelerating disease hypothesis generation. Nevertheless, large language models, which are typically trained on human-generated text, may not have the ability to determine the accuracy and appropriateness of the input data. As a result, they could inadvertently perpetuate human biases and preconceived notions. Moreover, given that these models rely heavily on published data, they may have limited potential to identify genuinely novel targets. Therefore, it is important to acknowledge these limitations and to complement their use with other models to ensure the discovery of truly novel and pertinent targets.

#### The use of AI-generated synthetic data for target identification

'Synthetic data' refers to artificially generated data that mimic real-world patterns and characteristics. By leveraging AI algorithms, synthetic data can be created to simulate various biological scenarios, thus enabling researchers to explore and analyze a broader range of possibilities [57–59]. This approach can be particularly valuable in therapeutic areas where experimental data are scarce or difficult to obtain. For example, in rare diseases or conditions where patient data are limited, AI can generate synthetic data based on existing knowledge and patterns. These synthetic data can then be used to train AI models and identify potential therapeutic targets that may have been overlooked [60]. Synthetic data can also be used to validate predictions made by AI algorithms, thus providing an additional layer of confidence in the target discovery process.

Furthermore, AI-generated synthetic data can help to address data imbalance or bias issues. In some therapeutic areas, particular patient populations may be under-represented in the available datasets, leading to challenges in target identification. AI can generate synthetic data representing these under-represented populations, allowing more comprehensive and inclusive analysis [61].

Although AI-generated synthetic data can offer advantages in exploring a broader range of possibilities and addressing data scarcity, it is essential to recognize its limitations. A model cannot simulate data containing complexities that the model is unaware of, and this limitation should be fully acknowledged [62]. Simulating under-represented populations, although tempting due to data scarcity, raises ethical concerns because collecting relevant data should be pursued whenever possible rather than relying solely on synthetic data [63,64]. Moreover, ensuring that the synthetic data accurately capture the intricate and nuanced aspects of real-world biological systems presents a significant challenge. Therefore, implementing robust validation and quality control measures becomes crucial to establish the reliability and relevance of the generated data [65].

To responsibly validate and control the quality of synthetic omic data, several options can be considered. First, comparative analyses can be performed to assess the similarity between the synthetic data and real-world data. This can involve statistical measures, such as comparing distributional characteristics, correlation patterns, or feature-level comparisons. In addition, benchmarking against known ground-truth data, where available, can help to evaluate the accuracy and performance of the synthetic data [66]. Another approach involves conducting functional analyses, such as focusing on the representation of particular cellular types in the synthetic dataset in the case of single-cell data, to determine whether the synthetic data captures biological knowledge and exhibits coherent functional relationships [67]. Finally, involving domain experts and conducting rigorous peer review can provide valuable insights and ensure the appropriateness and relevance of the synthetic data for target identification [59]. These validation and quality control measures, although challenging, can contribute to establishing confidence in the use of synthetic omic data in research and drug target discovery.

### Target selection criteria

The criteria used to select drug targets can greatly impact on the success of drug development (Figure 3). Causality represents a crucial criterion for selecting drug targets. Understanding the causal mechanisms behind a disease can help researchers to identify driver genes and key pathways that have the greatest potential for effective disease treatment [68]. Apart from experimental methods, a common computational approach to infer causal relationships between targets and diseases is network-based analysis, which involves the construction of biological networks that capture the relationships between different genes, proteins, drugs, and other molecular entities [69]. These networks can be used to identify potential targets that might have a causal involvement in a disease based on their centrality and connectivity within the network. The growing interest in AI and computational biology has led to a need for the development of machine learning methods that can be utilized for causal inference in biological networks [70]. In this regard, the adaptation of classification algorithms for causal discovery marks the emergence of causal inference models in biomedical research [71–73].

Another important consideration is the druggability of a target – the ability of a target to be modulated by a drug molecule. Factors that affect druggability include **therapeutic modality**, protein localization, class, and structure availability. For instance, small-molecule drugs are typically used for targets with well-defined binding pockets (e.g., kinases), whereas protein-based therapies are more suitable for targets that are difficult to tackle with small molecules. Structural information on drug targets is helpful for drug design and optimization with AI-based predictions, such as AlphaFold [74], thus expanding protein structure coverage. Target toxicity must also be considered by assessing the cellular processes, gene essentiality, and tissue specificity involved.

### Trade-off between high-confidence and novel targets

Novelty is another crucial factor in target selection in addition to causality, druggability, and toxicity. Text-based evidence can be used to assess novelty and confidence of a given target. Through scrutinizing the relationship between approved drugs, molecular targets, and therapeutic indications, Santos *et al.* revealed that high-confidence targets (or 'privileged' target families) accounted for the majority of approved drugs, whereas drugs tackling novel first-in-class targets represented only a small proportion, although this is increasing, especially in the field of oncology [75]. Striking a balance between novelty and confidence is essential for target selection. AI-powered **natural language processing** methodologies can aid this target selection process by extracting supporting evidence connecting a potential target to an indication based on huge amounts of data involving scientific publications, grants, and clinical trials, and this can provide a quantifiable scale for the novelty and confidence of targets in the context of the disease and enable flexible target-hunting workflows [76]. In addition, tools have been developed to quantify target novelty and confidence. TIN-X is an example that uses text-mining data processed from the scientific literature to quantify target novelty and confidence by providing two bibliometric indices, namely the 'novelty index' that represents the scarcity of target-associated publications, and the 'importance index' that assesses the strength of the association between a given target–disorder pair [77]. Furthermore, AI could facilitate drug repurposing by connecting a high-confidence target with known drugs to new disorders where the drugs have not been investigated, enabling cost-effective and time-saving drug discovery for both common and rare diseases [78].

### AI-identified targets validated in experiments

Target validation using cell and animal models is a crucial step in target discovery to reduce the project attrition rate and the cost of drug development in the pharmaceutical industry (Box 1). An increasing number of AI-identified targets are being successfully validated. For example, 28 AI-proposed targets for ALS treatment were validated in an ALS-mimicking *Drosophila* model,



### Box 1. Advances in target validation

Target validation using both cell and animal models is crucial to confirm the modulatory effects of the proposed target on disease development. Although 2D cell culture and rodent models are the prevailing tools for target validation, the difficulty of system establishment and the lack of complexity or recapitulation of human development limit their power as highly representative models. Organoids – 3D cell models derived from either **induced pluripotent stem cells (iPSCs)** or adult stem cells (ASCs) – have arisen as a promising technique for both disease research and drug testing by allowing the capture of tissue architecture and cellular microenvironment *in vitro* [84]. Taking advantage of their self-organizing ability, organoids are able to mimic actual organ development, and have been successfully established for multiple human organs (e.g., intestine, stomach, lung, liver, kidney, and brain) to explore the pathogenic mechanisms of various diseases [85–87]. Furthermore, because patient-derived organoids can retain the genetic, histopathological, and therapeutic response phenotypes of the primary disease tissue, these models have made their way into identifying personalized therapeutic regimens and drug efficacy testing [88,89]. In colorectal cancer, patient-derived colon organoids served as an effective tool to evaluate the efficacy of CAR-T cell therapy [90].

In both industrial and clinical laboratories there is a tendency to adopt automation to streamline experiments, data collection, and data analysis. With recent breakthroughs in bioengineering and machine learning, laboratory automation can greatly improve work efficiency and reproducibility by increasing data generation rate, reducing human technical variation, and avoiding contaminant exposure [91,92]. The development and commercialization rate of novel therapeutic interventions can also be enhanced by automation. For example, Insilico Medicine have launched an AI-driven robotic laboratory that is an interconnected expansion of their end-to-end AI drug discovery platform<sup>ix</sup>. Despite several remaining obstacles, the progressive integration of automation will revolutionize the laboratory environment to maximize research success.

revealing eight unreported targets whose suppression strongly rescues eye neurodegeneration [53]. In addition, in the same therapeutic area, Zhang *et al.* developed a machine learning-based method to identify *KANK1* as a novel gene linked to ALS and validated the neurotoxic effects of *KANK1* mutations reproduced by CRISPR–Cas9 in human neurons [79]. Inhibition of HDAC6 was identified as a cardioprotective strategy by deep learning, and was validated via a BAG3 cardiomyocyte-knockout mouse model of dilated cardiomyopathy [80]. CDK20 was identified as a target for the treatment of HCC using deep learning-based methods, and a highly potent small-molecule inhibitor designed by generative AI demonstrated selective antiproliferation activity in an HCC cell line [81]. Furthermore, Zeng *et al.* developed deepDTnet based on 15 heterogeneous types of chemical, genomic, phenotypic, and cellular networks to facilitate *in silico* identification of molecular targets for known drugs [82]. One of the identified drugs specifically targeting human ROR- $\gamma$ t shows therapeutic effects in a mouse model of multiple sclerosis.

### Concluding remarks and future perspectives

Target discovery is a crucial initial step in the modern drug discovery pipeline. Given that only a small proportion of the potentially druggable targets in humans have been identified, there is a pressing need for effective target discovery methods. The growing number of AI-identified targets being validated in experiments highlights the benefits of incorporating AI algorithms into target identification to enhance the efficiency of novel target discovery and the development of new therapeutics.

One area where AI is expected to make significant contributions is in tackling complex diseases. Diseases such as cancer, neurodegenerative disorders, and autoimmune conditions often involve intricate molecular mechanisms that are challenging to unravel. AI-driven target discovery methods can help to uncover novel targets and pathways underlying these diseases, paving the way for the development of more effective treatments.

Moreover, unexpected infectious disease outbreaks pose a constant threat to global health. The rapid identification of potential drug targets and the development of antiviral therapies are crucial for combating emerging pathogens<sup>v</sup>. By analyzing genomic data, AI algorithms can aid the identification of essential viral proteins or host factors that can be targeted to inhibit viral replication, thus providing valuable insights for the development of antiviral drugs [83].

### Outstanding questions

Can AI algorithms accurately predict target validation results and adverse effects, as well as druggability, specificity, off-target effects, and potential interactions with other drugs, for potential targets across different test systems (cell lines, animals, and humans)?

How can AI-driven target discovery approaches be validated, benchmarked against traditional experimental methods, and also effectively incorporate domain knowledge and expert insights to ensure reliability, reproducibility, and enhanced target identification and validation?

How can AI algorithms uncover the full mechanism of action at selected targets, consider the heterogeneity and variability of diseases including individual variations, and leverage this understanding to optimize combination therapies, leading to the identification of synergistic drug–target combinations for improved treatment outcomes?

How can we validate the reliability and robustness of predictions and discoveries based on synthetic AI-generated data, and how does it compare to experimental validation using real-world data?

AI also has the potential to revolutionize the discovery of efficient combinations of therapeutic targets and mechanisms. Complex diseases often involve multiple molecular pathways and interplay among various biological factors. AI algorithms can analyze diverse datasets, including genomic data, patient records, and synthetic lethality, to identify synergistic combinations of targets and mechanisms that may offer enhanced therapeutic effects. This approach can potentially transform treatment strategies, particularly in diseases where monotherapies have shown limited effectiveness.

Furthermore, the integration of AI with fully automated robotic laboratories offers the potential for high-throughput target validation and screening. Automated experiments, coupled with AI-driven data analysis, can expedite the validation of predicted targets, enabling researchers to assess their therapeutic potential quickly. This combination of AI and automation has the potential to revolutionize the drug discovery process and significantly reduce the time and cost required for target identification and validation.

Despite the tremendous progress made in AI-driven target discovery, several outstanding questions and challenges remain (see Outstanding questions). Ethical considerations, data privacy, and regulatory frameworks are crucial aspects that must be addressed to ensure responsible and ethical deployment of AI in drug development. Furthermore, the interpretability and explainability of AI algorithms are essential for gaining trust and acceptance from the scientific and medical communities. It is pertinent to note that, although AI has demonstrated potential in expediting the early stages of drug discovery such as target identification and lead optimization, it cannot significantly shorten the time required for clinical trials during drug development. This is because of the long period of time spent on ethical and regulatory approval, patient recruitment, duration of treatment, and data analysis, irrespective of whether the drug was developed by AI or not.

In summary, AI has emerged as a powerful tool in target discovery and drug development, and is revolutionizing how we identify novel drug targets and repurpose existing drugs. With the continued advancements in AI technology and the collaborative efforts of researchers, we can look forward to a future where AI plays an indispensable role in accelerating the development of safe and effective therapeutics for a wide range of diseases, ultimately improving human health and well-being.

### Acknowledgments

We thank Drs Feng Ren and Alex Aliper for expert advice, and Drs Xi Long and Bonnie Hei Man Liu for graphical illustrations and literature review.

### Declaration of interests

F.W.P., I.V.O., and A.Z. are employees of Insilico Medicine Hong Kong Ltd.

### Resources

<sup>i</sup><https://ftloscience.com/process-costs-drug-development/>

<sup>ii</sup>[www.fiercebitech.com/medtech/breaking-big-pharma-s-ai-barrier-insilico-medicine-uncovers-novel-target-new-drug-for](http://www.fiercebitech.com/medtech/breaking-big-pharma-s-ai-barrier-insilico-medicine-uncovers-novel-target-new-drug-for)

<sup>iii</sup>[www.nature.com/articles/d43747-021-00045-7](http://www.nature.com/articles/d43747-021-00045-7)

<sup>iv</sup>[www.eurekalert.org/news-releases/982543](http://www.eurekalert.org/news-releases/982543)

<sup>v</sup>[www.prnewswire.com/news-releases/insilico-medicine-announces-novel-3cl-protease-inhibitor-preclinical-candidate-for-covid-19-treatment-301553766.html](http://www.prnewswire.com/news-releases/insilico-medicine-announces-novel-3cl-protease-inhibitor-preclinical-candidate-for-covid-19-treatment-301553766.html)

<sup>vi</sup>[www.exscientia.ai/dsp-1181](http://www.exscientia.ai/dsp-1181)

<sup>vii</sup><https://investors.exscientia.ai/press-releases/press-release-details/2021/exscientia-announces-second-molecule-created-using-ai-from-sumitomo-dainippon-pharma-collaboration-to-enter-phase-1-clinical-trial/Default.aspx>

<sup>viii</sup><https://investors.exscientia.ai/press-releases/press-release-details/2023/Exscientia-Announces-First-in-Human-Study-for-Bristol-Myers-Squibb-In-Licensed-PKC-Theta-Inhibitor-EXS4318/default.aspx>

<sup>ix</sup>[www.globenewswire.com/news-release/2023/01/05/2583816/0/en/Insilico-Medicine-launches-6th-generation-Intelligent-Robotics-Lab-to-further-accelerate-its-AI-driven-drug-discovery.html](http://www.globenewswire.com/news-release/2023/01/05/2583816/0/en/Insilico-Medicine-launches-6th-generation-Intelligent-Robotics-Lab-to-further-accelerate-its-AI-driven-drug-discovery.html)

## References

- Hinkson, I.V. *et al.* (2020) Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. *Front. Pharmacol.* 11, 770
- Zhou, Y. *et al.* (2022) Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Res.* 50, D1398–D1407
- Kana, O. and Brylinski, M. (2019) Elucidating the druggability of the human proteome with eFindSite. *J. Comput. Aided Mol. Des.* 33, 509–519
- Finan, C. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9, eaag1166
- Sun, D. *et al.* (2022) Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* 12, 3049–3062
- Shiraiwa, K. *et al.* (2020) Chemical tools for endogenous protein labeling and profiling. *Cell Chem. Biol.* 27, 970–985
- van der Zouwen, A.J. and Witte, M.D. (2021) Modular approaches to synthesize activity- and affinity-based chemical probes. *Front. Chem.* 9, 644811
- Ong, S.E. and Mann, M. (2006) A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* 1, 2650–2660
- Jin, J. *et al.* (2023) SIRT3-dependent deacetylation of cyclin E2 prevents hepatocellular carcinoma growth. *EMBO Rep.* 24, e56052
- Li, X. *et al.* (2023) Deacetylation induced nuclear condensation of HP1 $\gamma$  promotes multiple myeloma drug resistance. *Nat. Commun.* 14, 1290
- Wang, Y. *et al.* (2022) DUT enhances drug resistance to proteasome inhibitors via promoting mitochondrial function in multiple myeloma. *Carcinogenesis* 43, 1030–1038
- Montero-Calle, A. *et al.* (2023) In-depth quantitative proteomics analysis revealed C1GALT1 depletion in ECC-1 cells mimics an aggressive endometrial cancer phenotype observed in cancer patients with low C1GALT1 expression. *Cell Oncol. (Dordr)* 46, 697–715
- Kortum, B. *et al.* (2022) Combinatorial treatment with statins and niclosamide prevents CRC dissemination by unhinging the MAOCC1– $\beta$ -catenin–S100A4 axis of metastasis. *Oncogene* 41, 4446–4458
- Qi, T.F. *et al.* (2023) Parallel-reaction monitoring revealed altered expression of a number of epitranscriptomic reader, writer, and eraser proteins accompanied with colorectal cancer metastasis. *Proteomics* 23, e2200059
- Nidhi, S. *et al.* (2021) Novel CRISPR–Cas systems: an updated review of the current achievements, applications, and future research perspectives. *Int. J. Mol. Sci.* 22, 3327
- Samelson, A.J. *et al.* (2022) BRD2 inhibition blocks SARS-CoV-2 infection by reducing transcription of the host cell receptor ACE2. *Nat. Cell Biol.* 24, 24–34
- Ramkumar, P. *et al.* (2020) CRISPR-based screens uncover determinants of immunotherapy response in multiple myeloma. *Blood Adv.* 4, 2899–2911
- Chakraborty, S. *et al.* (2018) Onco-Multi-OMICS approach: a new frontier in cancer research. *Biomed. Res. Int.* 2018, 9836256
- Nurk, S. *et al.* (2022) The complete sequence of a human genome. *Science* 376, 44–53
- Buniello, A. *et al.* (2019) The NHGRI–EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012
- Einarsson, G.G. *et al.* (2021) Extended-culture and culture-independent molecular analysis of the airway microbiota in cystic fibrosis following CFTR modulation with ivacaftor. *J. Cyst. Fibros.* 20, 747–753
- Sewell, G.W. and Kaser, A. (2022) Interleukin-23 in the pathogenesis of inflammatory bowel disease and implications for therapeutic intervention. *J. Crohns Colitis* 16, ii3–ii19
- Deelen, J. *et al.* (2019) Publisher correction: a meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* 10, 3669
- Namba, S. *et al.* (2022) A practical guideline of genomics-driven drug discovery in the era of global biobank meta-analysis. *Cell Genom.* 2, 100190
- Abell, N.S. *et al.* (2022) Multiple causal variants underlie genetic associations in humans. *Science* 375, 1247–1254
- Assum, I. *et al.* (2022) Tissue-specific multi-omics analysis of atrial fibrillation. *Nat. Commun.* 13, 441
- Suhre, K. *et al.* (2017) Erratum: connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* 8, 14357
- Yin, X. *et al.* (2022) Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. *Am. J. Hum. Genet.* 109, 1727–1741
- Mountjoy, E. *et al.* (2021) An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* 53, 1527–1533
- Na, F. *et al.* (2022) KMT2C deficiency promotes small cell lung cancer metastasis through DNMT3A-mediated epigenetic reprogramming. *Nat. Can.* 3, 753–767
- Gulfidan, G. *et al.* (2022) Systems biomarkers for papillary thyroid cancer prognosis and treatment through multi-omics networks. *Arch. Biochem. Biophys.* 715, 109085
- Lu, J. *et al.* (2021) Multi-omics analysis of fatty acid metabolism in thyroid carcinoma. *Front. Oncol.* 11, 737127
- Raivola, J. *et al.* (2022) Multiomics characterization implicates PTK7 in ovarian cancer EMT and cell plasticity and offers strategies for therapeutic intervention. *Cell Death Dis.* 13, 714
- Pinero, J. *et al.* (2018) Network, transcriptomic and genomic features differentiate genes relevant for drug response. *Front. Genet.* 9, 412
- Wolber, G. *et al.* (2008) Molecule–pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* 13, 23–29
- Lee, A. *et al.* (2016) Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov.* 11, 707–715
- Nettles, J.H. *et al.* (2006) Bridging chemical and biological space: 'target fishing' using 2D and 3D molecular descriptors. *J. Med. Chem.* 49, 6802–6810
- Lo, Y.C. *et al.* (2016) 3D chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS Chem. Biol.* 11, 2244–2253
- Vamathevan, J. *et al.* (2019) Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477
- Mamoshina, P. *et al.* (2018) Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* 9, 242
- Zhavoronkov, A. *et al.* (2019) Deep biomarkers of aging and longevity: from research to applications. *Aging (Albany NY)* 11, 10771–10780
- Muslu, O. *et al.* (2022) GuiltyTargets: prioritization of novel therapeutic targets with network representation learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 491–500
- Liu, R. *et al.* (2021) A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nat. Mach. Intell.* 3, 68–75
- Zhavoronkov, A. *et al.* (2019) Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040

45. Ivanenkov, Y.A. *et al.* (2023) Chemistry42: an AI-driven platform for molecular design and optimization. *J. Chem. Inf. Model.* 63, 695–701
46. Obrezanova, O. (2023) Artificial intelligence for compound pharmacokinetics prediction. *Curr. Opin. Struct. Biol.* 79, 102546
47. Chen, R. *et al.* (2018) Machine learning for drug–target interaction prediction. *Molecules* 23, 2208
48. Ye, Q. *et al.* (2021) A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nat. Commun.* 12, 6775
49. Kavalci, E. and Hartshorn, A. (2023) Improving clinical trial design using interpretable machine learning based prediction of early trial termination. *Sci. Rep.* 13, 121
50. McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133
51. Zhavoronkov, A. *et al.* (2019) Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Res. Rev.* 49, 49–66
52. Aliper, A. *et al.* (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530
53. Pun, F.W. *et al.* (2022) Identification of therapeutic targets for amyotrophic lateral sclerosis using pandaomics – an AI-enabled biological target discovery platform. *Front. Aging Neurosci.* 14, 914017
54. Fabris, F. *et al.* (2020) Using deep learning to associate human genes with age-related diseases. *Bioinformatics* 36, 2202–2208
55. West, M.D. *et al.* (2018) Use of deep neural network ensembles to identify embryonic-fetal transition markers: repression of COX7A1 in embryonic and cancer cells. *Oncotarget* 9, 7796–7811
56. Luo, R. *et al.* (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* 23, bbac409
57. Shayakhmetov, R. *et al.* (2020) Molecular generation for desired transcriptome changes with adversarial autoencoders. *Front. Pharmacol.* 11, 269
58. Vinas, R. *et al.* (2022) Adversarial generation of gene expression data. *Bioinformatics* 38, 730–737
59. Beaulieu-Jones, B.K. *et al.* (2019) Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* 12, e005122
60. Song, J. *et al.* (2021) The discovery of new drug–target interactions for breast cancer treatment. *Molecules* 26, 7474
61. Chawla, N.V. *et al.* (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357
62. Achuthan, S. *et al.* (2022) Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. *J. Biosci.* 47, 43
63. Howe, E.G., III and Elenberg, F. (2020) Ethical challenges posed by big data. *Innov. Clin. Neurosci.* 17, 24–30
64. Bhanot, K. *et al.* (2021) The problem of fairness in synthetic healthcare data. *Entropy (Basel)* 23, 1165
65. Rajotte, J.F. *et al.* (2022) Synthetic data as an enabler for machine learning applications in medicine. *iScience* 25, 105331
66. El Emam, K. *et al.* (2022) Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med. Inform.* 10, e35734
67. Treppner, M. *et al.* (2021) Synthetic single cell RNA sequencing data from small pilot studies using deep generative models. *Sci. Rep.* 11, 9403
68. Nogales, C. *et al.* (2022) Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.* 43, 136–150
69. Buphalmai, P. *et al.* (2021) Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun.* 12, 6306
70. Lecca, P. (2021) Machine learning for causal inference in biological networks: perspectives of this challenge. *Front. Bioinform.* 1, 746712
71. Cassan, O. *et al.* (2021) Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics* 22, 387
72. Gillani, Z. *et al.* (2014) CompareSVM: supervised, support vector machine (SVM) inference of gene regularity networks. *BMC Bioinformatics* 15, 395
73. Zhou, X. and Kosorok, M.R. (2017) Causal nearest neighbor rules for optimal treatment regimes. *ArXiv* Published online November 22, 2017. <https://arxiv.org/abs/1711.08451>
74. Varadi, M. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50, D439–D444
75. Santos, R. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16, 19–34
76. Vera, C.D. *et al.* (2022) Treating Duchenne muscular dystrophy: the promise of stem cells, artificial intelligence, and multi-omics. *Front. Cardiovasc. Med.* 9, 851491
77. Cannon, D.C. *et al.* (2017) TIN-X: target importance and novelty explorer. *Bioinformatics* 33, 2601–2603
78. Pushpakom, S. *et al.* (2019) Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* 18, 41–58
79. Zhang, S. *et al.* (2022) Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron* 110, 992–1008
80. Yang, J. *et al.* (2022) Phenotypic screening with deep learning identifies HDAC6 inhibitors as cardioprotective in a BAG3 mouse model of dilated cardiomyopathy. *Sci. Transl. Med.* 14, eabl5654
81. Ren, F. *et al.* (2023) AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chem. Sci.* 14, 1443–1452
82. Zeng, X. *et al.* (2020) Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797
83. Ong, E. *et al.* (2020) COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. *Front. Immunol.* 11, 1581
84. Jensen, C. and Teng, Y. (2020) Is it time to start transitioning from 2D to 3D cell culture? *Front. Mol. Biosci.* 7, 33
85. Fan, W. *et al.* (2022) Applications of brain organoids for infectious diseases. *J. Mol. Biol.* 434, 167243
86. Sidhaye, J. and Knoblich, J.A. (2021) Brain organoids: an ensemble of bioassays to investigate human neurodevelopment and disease. *Cell Death Differ.* 28, 52–67
87. Angus, H.C.K. *et al.* (2019) Intestinal organoids as a tool for inflammatory bowel disease research. *Front. Med. (Lausanne)* 6, 334
88. Wensink, G.E. *et al.* (2021) Patient-derived organoids as a predictive biomarker for treatment response in cancer patients. *NPJ Precis. Oncol.* 5, 30
89. Berkers, G. *et al.* (2019) Rectal organoids enable personalized treatment of cystic fibrosis. *Cell Rep.* 26, 1701–1708
90. Schnalzger, T.E. *et al.* (2019) 3D model for CAR-mediated cytotoxicity using patient-derived colorectal cancer organoids. *EMBO J.* 38, e100928
91. Burger, B. *et al.* (2020) A mobile robotic chemist. *Nature* 583, 237–241
92. Crone, M.A. *et al.* (2020) A role for biofoundries in rapid development and validation of automated SARS-CoV-2 clinical diagnostics. *Nat. Commun.* 11, 4464

## Review

# Deep generative molecular design reshapes drug discovery

Xiangxiang Zeng,<sup>1</sup> Fei Wang,<sup>2</sup> Yuan Luo,<sup>3</sup> Seung-gu Kang,<sup>4</sup> Jian Tang,<sup>5</sup> Felice C. Lightstone,<sup>6</sup> Evandro F. Fang,<sup>7,8</sup> Wendy Cornell,<sup>4</sup> Ruth Nussinov,<sup>9,10</sup> and Feixiong Cheng<sup>11,12,13,\*</sup>

<sup>1</sup>College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan 410082, P.R. China

<sup>2</sup>Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY 10065, USA

<sup>3</sup>Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

<sup>4</sup>Healthcare & Life Sciences Research, IBM TJ Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA

<sup>5</sup>Mila-Quebec Institute for Learning Algorithms and CIFAR AI Research Chair, HEC Montreal, Montréal, QC H3T 2A7, Canada

<sup>6</sup>Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Lab, Livermore, CA 94550, USA

<sup>7</sup>Department of Clinical Molecular Biology, University of Oslo and Akershus University Hospital, 1478 Lørenskog, Oslo, Norway

<sup>8</sup>The Norwegian Centre on Healthy Ageing (NO-Age), Oslo, Norway

<sup>9</sup>Computational Structural Biology Section, Frederick National Laboratory for Cancer Research in the Laboratory of Cancer Immunometabolism, National Cancer Institute, Frederick, MD 21702, USA

<sup>10</sup>Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

<sup>11</sup>Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA

<sup>12</sup>Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA

<sup>13</sup>Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA

\*Correspondence: [chengf@ccf.org](mailto:chengf@ccf.org)

<https://doi.org/10.1016/j.xcrm.2022.100794>

## SUMMARY

Recent advances and accomplishments of artificial intelligence (AI) and deep generative models have established their usefulness in medicinal applications, especially in drug discovery and development. To correctly apply AI, the developer and user face questions such as which protocols to consider, which factors to scrutinize, and how the deep generative models can integrate the relevant disciplines. This review summarizes classical and newly developed AI approaches, providing an updated and accessible guide to the broad computational drug discovery and development community. We introduce deep generative models from different standpoints and describe the theoretical frameworks for representing chemical and biological structures and their applications. We discuss the data and technical challenges and highlight future directions of multimodal deep generative models for accelerating drug discovery.

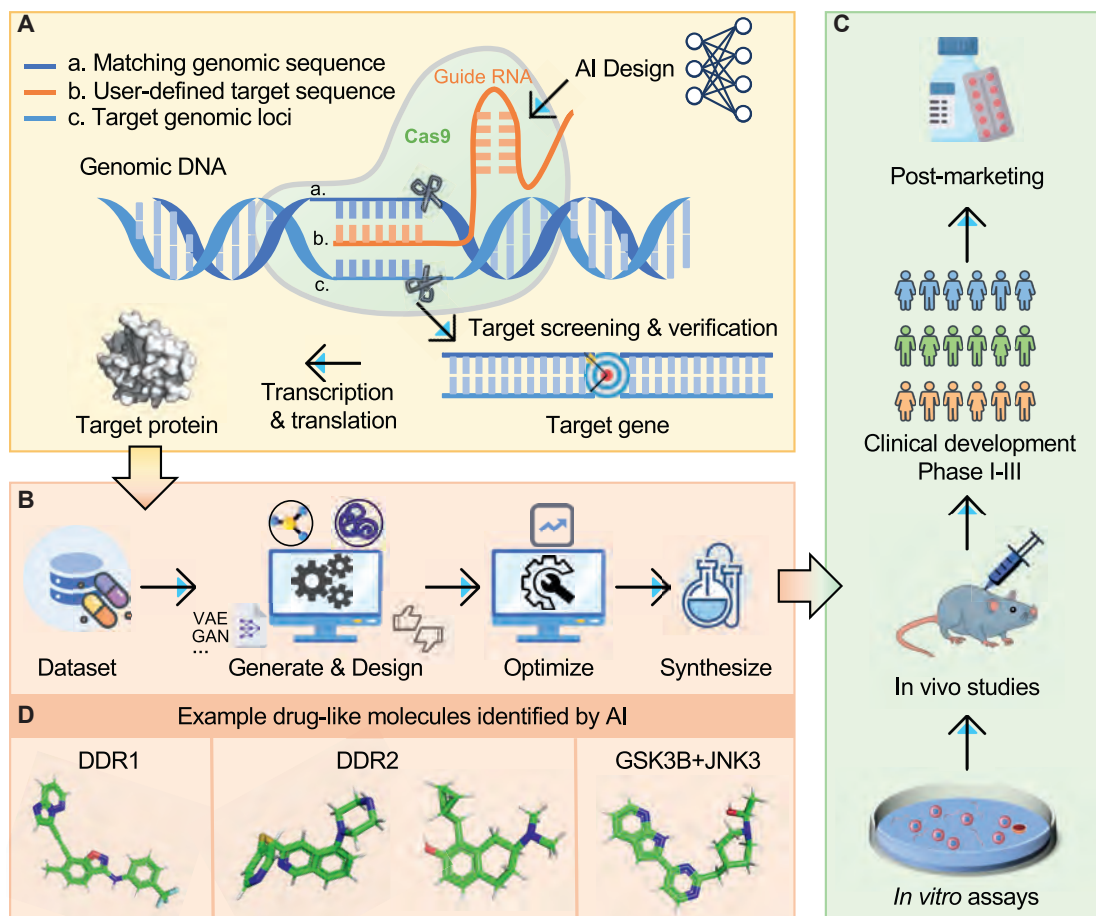
## INTRODUCTION: DEEP GENERATIVE MODELS IN DRUG DISCOVERY

A recent study estimates that pharmaceutical companies spent \$2.6 billion in 2015 for the development of new, US Food and Drug Administration-approved drugs, up from \$802 million in 2003.<sup>1</sup> Although more direct costs are incurred during clinical trials, since the preclinical investment comes earlier the capitalized costs of the two stages are roughly equal. Recent advances in computational sciences and technologies capture the requisites and urgencies and provide a set of potentially promising approaches. Among these, the developers can select the right artificial intelligence (AI) to target the problem at hand, in particular deep generative models, appropriate protocol, and factors. Collectively, they map paths that integrate biology, chemistry, computational science, pharmacology, and disease treatments.

The rapid growth in computing power, amount of data, and advanced algorithms has led to breakthroughs in AI for drug discovery,<sup>2</sup> especially in the application of deep generative models.<sup>3–5</sup> The models have emerged as high potential tools to transform the design, optimization, and synthesis of small molecules, and macromolecules (Figure 1). Applications of deep generative models have already delivered new partially optimized candidate leads, in some cases in less time typically required by conventional sequential approaches.<sup>6–10</sup> If applied on a large scale, deep generative modeling has the potential of boosting the development (R&D) process.

Deep generative models correspond to a theoretical framework for generating novel chemical and biological structures with desired properties using data structures, such as graphs and fingerprints, and operations, such as the flow of functional or experimental information. Creative deep generative models





**Figure 1. AI and deep generative model applications in the drug discovery pipeline**

Several successful applications of AI and deep generative models in various stages of the drug development pipeline: (A) AI-assisted target selection and validation, (B) molecular design, lead optimization, and chemical synthesis, (C) biological evaluation (*in vitro* and *in vivo*), clinical development, and post-marketing surveillance, and (D) several successful preclinical and clinical molecules identified by AI and deep generative models. DDR1, discoidin domain receptor 1; DDR2, discoidin domain receptor tyrosine kinase 2; GSK3B, glycogen synthase kinase 3 beta; JNK3, c-Jun N-terminal kinase 3.

can significantly promote algorithm development and application in drug discovery. In this “big data” era, deep generative models would offer a cutting-edge technology that could revolutionize an informatics view of biology, disease, and therapeutics. In this review, we describe classical and state-of-the-art deep generative models and their applications (Figure 1) in computational drug discovery and discuss limitations and challenges. Our aim is to provide an overview of current tools and techniques (the toolbox) of deep generative models in multiple applications on small-molecule and macromolecular systems.

## THE TOOLBOXES FOR DEEP GENERATIVE MODELS

Designing a novel drug is a complex undertaking that needs to satisfy pre-defined criteria for on-target potency, specificity relative to off-targets, physical properties, and other chemistry and biology measures. Traditional methods, which require chemists to select and validate candidate molecules experimentally from

a vast chemical space, are ineffective. Deep generative models have become popular because they can automatically generate new bioactive and synthesizable molecules in a time- and cost-effective way.

## Big biomedical datasets for drug discovery

We begin with a brief overview of several commonly used chemical and bioinformatics databases, which provide both labeled and unlabeled data to train, validate, and test deep generative models for the drug discovery community. Pharmaceutical companies have their in-house proprietary collections on the order of 2–3M compounds with associated data from past drug discovery quests. In the public domain, the ZINC database collected nearly 2 billion purchasable, commercially available, “drug-like” compounds for *in silico* screening.<sup>11</sup> Its massive size makes it also useful for learning molecular patterns for pre-training generative models. Bioactive molecules, such as those in the manually curated ChEMBL database, which approaches 1.5M of real bioactive molecules with every molecule having at least

one experimental bioactivity measurement,<sup>12</sup> are of particular interest. They can be used for training models to generate molecules with certain properties. The GDB-17 database<sup>13</sup> enumerates most organic molecules (166.4 billion) of up to 17 heavy atoms of C, N, O, S, and halogens. This includes many of the lower-molecular-weight small-molecule drugs as well as the smaller typical lead compounds. Ultra-large chemical databases,<sup>14</sup> such as Enamine (<https://enamine.net>) and REALdb,<sup>15</sup> contain billions of synthesizable compounds identified by chemoinformatics approaches and expert-system type rules. These ultra-large databases offer an opportunity to train models with broadened applicability. In addition to small-molecule resources, several macromolecular databases offer enriched data for generative model training in macromolecule design, such as the PDB.<sup>16</sup>

### Representation of compounds/molecules

The representation of molecules is important for generative models. There are three types of representations: (1) sequence based, (2) graph based, and (3) images (Figure 2). The unprecedented success of natural language processing (NLP) inspired the idea to describe molecules in symbols in a way analogous to human language. Semantics and grammars in biological structures bear a resemblance to human language; hence, molecules can be represented as sequences of characters. *De novo* small-molecule designs generally use simplified molecular input line entry systems (SMILES).<sup>17</sup> The sequence-based structure is generated by following the SMILES grammar rules encoded into vectors (Figure 2A). A more direct method to represent molecules is graph based.<sup>18</sup> In the graph representation, the atoms of a small molecule form a set of nodes and the bonds are regarded as edges (Figure 2B). For macromolecules, a contact map<sup>19</sup> is a graph that denotes the distance between any two amino acid residue pairs. Training graph-based models on a large number of nodes is expensive because the space complexity increases with the square of their number.<sup>20</sup> Compared with sequence-based approaches, graph-based representations are easy to implement as graph convolutional layers, and bond weights can be optimized in message-passing networks. Sequence-based representations are in general compact, memory-efficient, and easily searchable. However, both sequence-based and graph-based approaches cannot capture the 3D information of ligands or proteins in biologically meaningful ligand-protein interactions. The 3D conformation of a molecule captures the relative orientation of atoms<sup>21–24</sup> (Figure 2C). Several latest 3D representations were presented as well.<sup>25–27</sup> DEVELOP incorporate an existing graph-based deep generative model, De-Linker, along with a convolutional neural network to utilize 3D representations of molecules and target pharmacophores.<sup>28</sup> DeepLigBuilder is a graph-based generative model that utilizes 3D structural representation of ligand-receptor interactions for the end-to-end design of chemically and conformationally valid 3D molecules with drug-likeness properties.<sup>29</sup> Traditional image or 3D representation of proteins requires accurate 3D structural data from cryoelectron microscopy and crystallography, which is challenging to obtain. Recent

AI approaches, such as AlphaFold2, can provide massive protein 3D data to address these challenges.<sup>30</sup>

### Recurrent neural networks

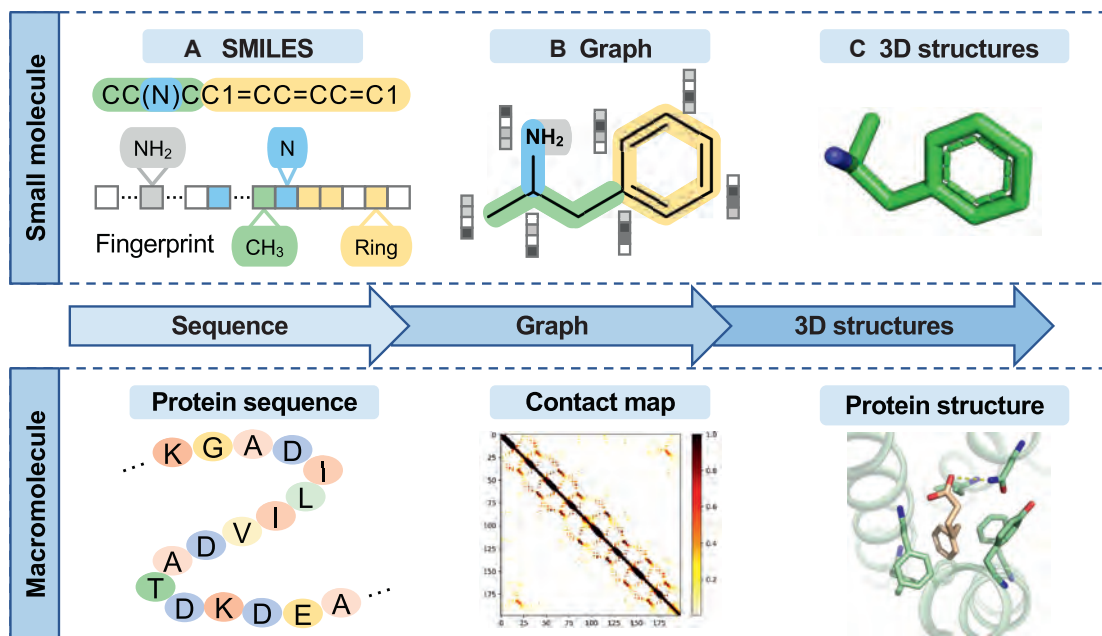
Recurrent neural networks (RNNs) are fundamental components of generative neural networks in processing human language. They are useful for modeling systems that have a sequential or time component and have been powerful in NLP automated computer code generation<sup>31</sup> and musical composition.<sup>32</sup> The language of molecules, such as SMILES, is similar to human language. Thus, it is natural to use RNNs for generating molecules based on sequential representation. As depicted in Figure 3A, SMILES (i.e., “c1cc ... c1”) can be generated by RNNs in the following way. RNNs receive the first character “c” and assign different probabilities to possible next characters: character “1” would receive a high probability and may be sampled as the next one. “1” is feedback input to RNNs. This process is repeated until the end token “\n” is generated. Long short-term memory (LSTM)<sup>33</sup> and gated recurrent unit (GRU)<sup>34</sup> introduce a gate mechanism to remember valuable input information for a long series of steps, lacking in traditional RNNs. Whether LSTM or GRU is preferable may depend on the specific application. LSTM cell can hold much longer history than GRU. However, additional parameters in LSTM may increase the risk of overfitting. RNNs with LSTM or GRU are among the most promising for the generation of *de novo* small molecules under the representation of SMILES.<sup>35</sup>

### Variational autoencoder

An autoencoder (AE) is constructed of two networks: (1) one (the encoder) is trained to map the input into a low-dimensional latent vector, and (2) the other (the decoder) to map the latent vector into the inputted data. The original AE creates a latent space by reproducing the input. To avoid overfitting and discontinuities in the original AE, variational AE (VAE) regularizes the latent space by replacing latent space points with distributions. In a pioneering work, VAE was employed for molecule generation, ushering in a new strategy in *de novo* drug design.<sup>10</sup> As shown in Figure 3C, the encoder is trained to map the molecules (e.g., SMILES) into a low-dimensional latent vector that is assumed to be sampled from a Gaussian distribution, and the decoder to map the latent vector into the inputted molecules (e.g., SMILES). The latent vectors are constrained to follow a probability distribution (usually Gaussian distribution) so that a molecule is represented as an explicit probability distribution over latent space. When the encoder and decoder are trained jointly, the output must reconstruct the training samples' probability distribution. Recently, learning disentangled representations for VAE has attracted increasing attention, where the main goal is to make each latent variable of the latent vector encode an independent property or factor of data.<sup>36</sup> If disentangled VAE is successfully introduced for molecular generation, a molecular property can be edited without changing other properties, by editing the latent variables associated with that property.

### Generative adversarial networks

The invention of generative adversarial networks (GANs)<sup>37</sup> started a flurry of generative models. Unlike VAE, GANs do not



**Figure 2. A diagram illustrating three molecular representation approaches**

Three molecular representation approaches include: (A) one-dimensional (1D) sequence-based representation; (B) graph-based representation; and (C) 3D representation for both small molecules and macromolecules (i.e., proteins). The value of contact map matrix is 1 if the distance is greater than a predetermined threshold, otherwise it is 0.

work with an explicit probability density function (Figure 3D), but provide an adversarial training framework composed of a generator and a discriminator. The discriminator trains a classification model aiming at maximizing the error rate of synthetic molecules from the generator, which resemble the real data. The generator and the discriminator are trained together in an adversarial, zero-sum game, until the discriminator model is fooled, meaning the generator network is generating plausible (i.e., realistic fake) molecules.

### Flow-based models

VAE and GAN do not explicitly model the real probability density function. VAE implicitly optimizes the log likelihood of the data by maximizing a lower bound on a likelihood function, whereas GAN avoids modeling the distribution but learns in an adversarial way to measure the difference between “valid molecules” and “synthetic molecules.” Deep flow-based models resolve the intractability issue of explicit density estimation by leveraging normalizing flow.<sup>38</sup> A normalizing flow is an invertible deterministic transformation between the raw data space and latent space (Figure 3B). For example, a recent method called MoFlow learns a chain of transformation to map valid molecules to their latent representations, and the reverse chain of transformation to map the latent representations to valid molecules.<sup>39</sup> One major limitation for the flow-based models is that they are time consuming due to the complex hyperparameter tuning processes. To take full advantage of the flow-based models, the molecular graphs must be transformed into continuous data by incorporating real-value noise into the molecular generation flow.

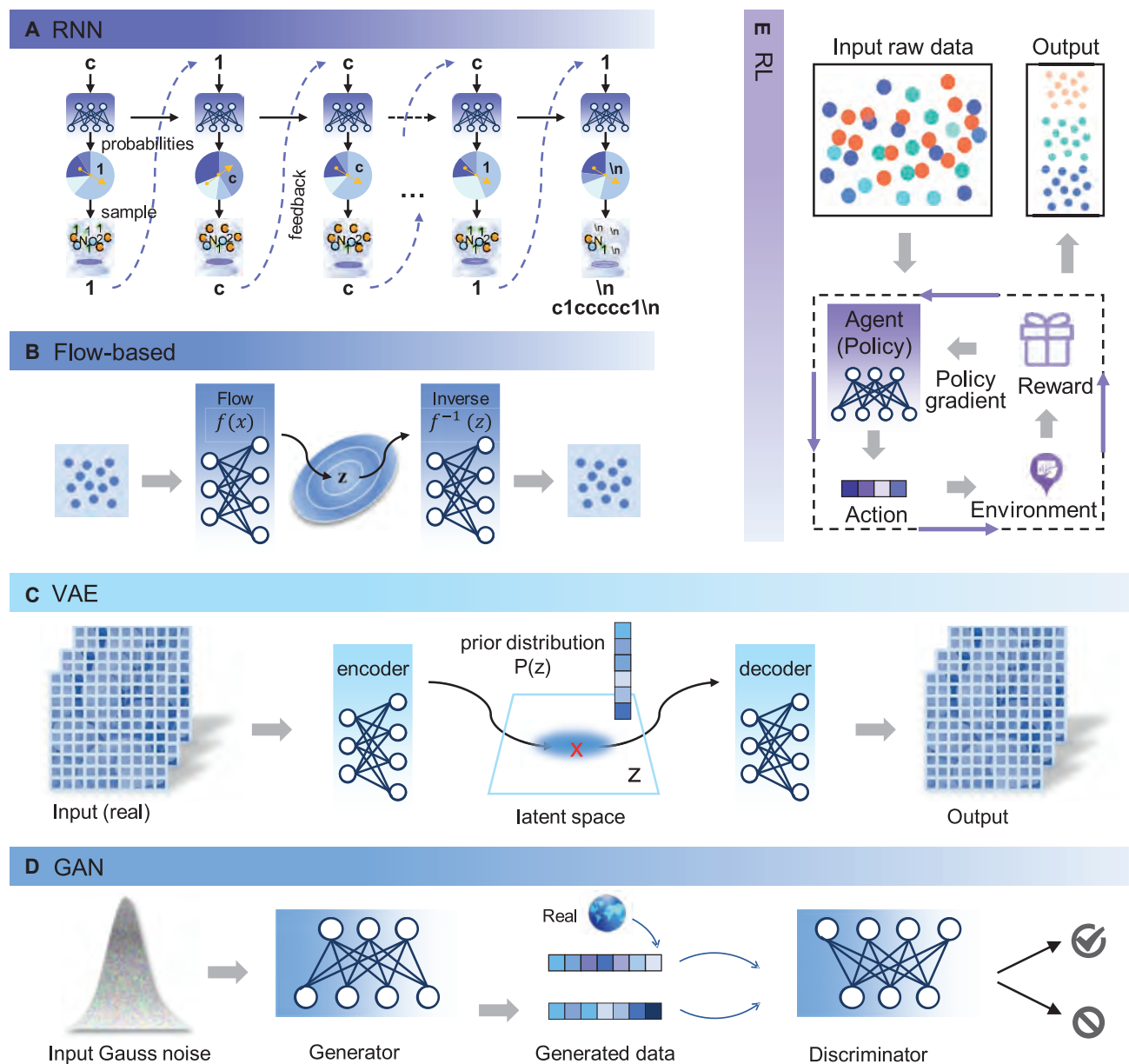
### Reinforcement learning

Deep RL has emerged as one of the most prominent toolboxes for optimizing an objective, especially with recent breakthroughs, such as AlphaGo.<sup>40</sup> The immensity of the chemical space is similar to Go’s enormous possible solution space; hence, RL is a potential method for exploring the chemical space by a dynamic decision process.<sup>41</sup> As depicted in Figure 3E, RL—consisting of an agent, a reward function, and environment—aims to optimize toward a user-directed target. The agent chooses the next action, and the reward function evaluates the quality of the actions according to the environment (domain-specific rules) and provides feedback to the agent. After the generative model is trained on a large and general set of molecules to learn the SMILES grammar, RL can be applied as a technique for fine-tuning of target properties, such as synthetic accessibility<sup>42</sup> and quantitative estimate of druglikeness,<sup>43</sup> which assesses physical properties. For example, policy gradient for forward synthesis (PGFS) (more below) was proposed to generate synthetically accessible molecules using RL.<sup>44</sup> For this, (1) the agent is a neural network; (2) the policy actions are chemical transformations executed by modifying a molecule by adding or removing atoms and bonds; and (3) the reward is synthetic accessibility.<sup>44</sup>

### APPLICATIONS IN SMALL-MOLECULE DRUG DESIGN

Conventional exploration, such as virtual screening,<sup>45,46</sup> needs to navigate a vast chemical space, posing time and cost challenges. *De novo* design, a technique of automatically generating molecules with desired properties from scratch, has benefited





**Figure 3. A diagram illustrating the theory framework of five deep generative models (A–E) in the drug discovery applications** RNN, recurrent neural networks; VAE, variational autoencoder; GAN, generative adversarial networks; RL, reinforcement learning.

from advances in deep generative models.<sup>47</sup> Here, we describe their applications toward various design purposes.

### Generating valid small molecules

As deep generative models for *de novo* small-molecule design were emerging, research initially focused on how to generate molecules with high validity, with a particular emphasis on the grammar and semantics of small molecules. In 2016, Gómez-Bombarelli et al. pioneered a data-driven method that generates molecules by mapping discrete high-dimensional chemical space to and from continuous latent space.<sup>10</sup> The model showed

that training VAE jointly with a molecular property prediction task and optimizing via a Gaussian process were promising. This paradigm promoted the development of *de novo* small-molecule design, even if the output included invalid molecules. Subsequently, inspired by the compiler theory where the syntax and semantics check is done via syntax-directed translation (SDT), Dai et al. incorporated SDT into VAE for constraining the decoder.<sup>48</sup> The proposed model (SD-VAE) can generate both syntactically and semantically valid molecules.<sup>48</sup>

Previous works achieved high validity by incorporating extra constraints. Inspired by fragment-based drug discovery, Jin

et al. proposed junction tree variational encoder (JT-VAE).<sup>49</sup> JT-VAE considers chemically valid substructures, such as aromatic rings as nodes in the graph structure. A molecular graph assembled by these nodes can maintain chemical validity without implementing additional chemical rules. JT-VAE reached 100% validity due to obeying the ground truth in chemistry by generating bioactive molecules from fragments. A new AE, the Wasserstein autoencoder character (cWAE),<sup>50</sup> incorporates adversarial training and has shown improved model accuracy. When applied to molecular design and trained on 1.6 billion compounds, compared with JT-VAE, cWAE produces an accurate generative model (the compound reconstruction error is reduced by over 80%).<sup>51</sup> MoFlow<sup>39</sup> generates a molecular graph in a one-shot manner that generates bonds and atoms by a flow-based model and then assembles them into a molecular graph. Instead, MolGrow<sup>52</sup> generates a molecular graph in an iterative manner, termed a hierarchical normalizing flow model via generating molecular graphs from a single-node graph by recursively splitting every node into two. Experimental results show that both MoFlow and MolGrow can generate 100% valid molecules.

### Generating molecules with drug-like properties

With the gradual maturity of generative models, molecular generative models have been aiming to find molecules with specific properties, not only focusing on their validity. Drug-like properties, such as biological activity and synthetic accessibility, are critical for the success of drug candidates. In 2020, a molecular GAN model<sup>53</sup> conditioned on gene expression signatures was shown to generate molecules with a high probability to induce a desired transcriptomic profile.

Generative tensorial reinforcement learning (GENTRL)<sup>54</sup> was designed to generate novel molecules that can inhibit DDR1 (discoidin domain receptor 1) by designing a reward function. The generated molecules were evaluated using *in vitro* and *in vivo* mouse assays to verify the binding affinity on DDR1 and the pre-clinical and pharmacokinetic properties. With a time frame of 46 days from target selection to partially validated molecule, GENTRL validated a promising outlook for accelerating drug discovery (Figure 1D). Notably, GENTRL leveraged a set of relevant information which is frequently available, such as crystal structure data and information related to active compounds. This model is not generalizable to cases where target-specific activity data are unavailable, and a model requiring less information could be more practical in such cases.

PGFS<sup>44</sup> was designed to generate molecules that can be feasibly synthesized. PGFS treats the molecular generation problem as a sequential decision process of selecting reactant molecules and reaction transformation in a linear synthetic sequence, where the choice of reactants is considered an action and synthetic accessibility a reward. PGFS has been validated in an *in-silico* proof-of-concept associated with three HIV targets.<sup>44</sup>

### Generating molecules with multi-objective drug-like properties

Generative models for *de novo* molecular generation are able to design molecules with multiple design constraints such as potency, safety, and desired metabolic profile. Molecules with such constraints will better meet the requirements of drug dis-

covery. RationaleRL<sup>55</sup> trained a graph-based RL model to complete a pre-selected molecular subgraph into an integral molecule with several desired co-existing properties, such as bioactivities toward multiple targets (e.g., GSK3 $\beta$  and JNK3; Figure 1D), quantitative estimate of drug-likeness, and synthetic accessibility. As part of multi-objective optimization, the predictiveness to drug-likeness has been significantly improved by combining individual classifiers and calculating their Bayesian errors. The difficulty lies in how to define and characterize non-drug-like molecules.<sup>56</sup>

### Generating better bioavailable molecules with optimization

Molecular optimization aims toward desired properties for a given starting molecule. This process is analogous to image-to-image translation (e.g., turn horses into zebras) in computer vision or style transfer in NLP. Jin et al. presented an optimization method inspired by style transfer.<sup>57</sup> Molecular optimization can be formulated as graph-to-graph translation via converting one molecular graph to another with better properties using the paired training sets.

Inspired by the image-to-image translation approach that CycleGAN<sup>58</sup> learned to translate an image from a source domain X to a target domain Y in the absence of paired examples, MolCycleGAN<sup>59</sup> was proposed and trained on two datasets with and without a desired property. The training framework consists of two GANs forming a cycle: (1) the first GAN is used to generate molecules with the desired property when the input is not equipped with the target property, and (2) the second network has the opposite input/output order. The objective of the model is to minimize the distance between the original molecules and the generated molecules of the second network.

### Capturing 3D information of ligand-protein interactions

In an attempt to bring 3D protein structure information directly into generative molecule creation rather than by post-generation docking, a high-quality target family sequence alignment was leveraged to identify binding site residues across the kinase family and train 1D string representation of the PaccMann model.<sup>60</sup> The quantitative structure-activity relationship (QSAR) model built with this reduced dataset outperformed the QSAR model built with the conventional full-sequence approach, and the molecules created with the generative model were likewise encouraging in terms of their similarity to validated kinase inhibitors.<sup>61</sup>

## APPLICATIONS IN MACROMOLECULAR DRUG DESIGN

In addition to designing small molecules, the application of AI has been extended to the design of medicinal macromolecules, such as designing antimicrobial peptides (AMPs), therapeutic proteins, and CRISPR-Cas9 systems design and optimization, as detailed below.

### AMP generation

The emergence of antibiotic-resistant bacteria led to nearly 1 million deaths worldwide each year from bacterial infections that cannot be treated with ordinary antibiotics.<sup>62</sup> AMPs increase the repertoire and deep generative models are a promising way

of designing them. Das et al. augmented a variant of VAE (Wasserstein Autoencoder)<sup>63</sup> with molecular dynamics information to generate AMPs with broad-spectrum potency and low toxicity.<sup>64</sup> For a controlled sequence generation, linear binary classifiers conditional latent (attribute) space sampling (CLaSS) for attribute prediction was trained on the latent space, and then rejected sampling was utilized for screening the molecules of interest. CLaSS can be trained for binary classification of antimicrobial function, broad-spectrum efficacy, presence of secondary structures, and toxicity at the same time. Within 48 days, two new antimicrobial peptides with high potency against Gram+ and Gram- bacteria were synthesized and tested *in vitro* and in mice. Both resulted in low resistance in *Escherichia coli* and low toxicity. Another example of antibiotic discovery emerged from combining the message-passing approach and experimental assays to predict the growth inhibition of *E. coli* followed by screening an existing compound library to identify molecules with antimicrobial activity and different structures from known antibiotics.<sup>9</sup> In the message-passing approach, the processors execute a task independently and communicate data between them by exchanging messages.

### Therapeutic protein generation

*De novo* protein design plays important roles in protein therapies. For instance, a *de novo* design strategy was proposed to produce rapidly and accurately decoy proteins by replicating the protein interface of human angiotensin I-converting enzyme 2 (hACE2) for a potential treatment of coronavirus disease 2019 (COVID-19).<sup>65</sup> Deep generative models can also be used to design protein therapies by modeling the spatial properties of the amino acid sequence. ProteinGAN,<sup>66</sup> which incorporated a self-attention mechanism into GAN and learned the evolutionary relationships of protein sequences, was a generalizable framework to generate protein sequences with specific functions. About 24% of the generated sequences were soluble and showed activity comparable with the wild types, including some highly mutated sequences. The generated sequences include 119 novel structural sequence motifs, not present in the training dataset, showcasing *de novo* generation of functional proteins for therapeutic development.

### CRISPR-Cas9 systems design and optimization

The CRISPR-Cas9 system, consisting of a Cas9 nuclease and a guide RNA (gRNA), is a technology for genome editing and a tool to identify targets in drug discovery (Figure 1A). Based on the principle of complementary base pairing, gRNA guides Cas protein localization to the genome and CRISPR KO (knockout). CRISPRi (interference) and CRISPRa (activation) technologies then determine whether the candidate genes are the key to disease and thus a therapeutic target. The selection of gRNA sequences affects knockout efficacy and is essential for target identification. Recent studies have demonstrated the power of deep learning algorithms, such as CNNs and RNNs, to design and optimize CRISPR-Cas9 systems. Recently, Chuai et al. proposed a design tool called DeepCRISPR for gRNA with high sensitivity and specificity, which adopts a combination of unsupervised and supervised CNNs to learn the representations of gRNAs.<sup>67</sup> DeepCRISPR can predict on-target knockout efficacy

and off-target profile in the same framework. In addition, it automatically detects important features of optimized gRNAs to promote effective CRISPR design. SpCas9 genome editing tools<sup>68</sup> can address the off-target issue. A DeepHF model, which combined RNNs with the secondary structure, GC content, and thermodynamics features was developed, but could not be automatically obtained by RNNs.<sup>69</sup> Although deep learning models have conveniently facilitated CRISPR-Cas9 systems design, these data-driven approaches are subject to the problems of data heterogeneity, sparsity, and imbalance.<sup>67</sup> CRISPR-Cas9 systems design can be further optimized using advanced algorithms with higher-quality data.

### OUTSTANDING QUESTIONS, PERSPECTIVE, AND FUTURE DIRECTION

Despite the enthusiasm for AI-enabled drug discovery, questions and challenges abound. For decades, translational science has been facing the challenge of how to translate research findings into a novel, more effective medicine.<sup>70</sup> In fact, the “ultimate goal of the translational challenge is to eliminate the Valley of Death, through scientific understanding and innovation.”<sup>71</sup> Most machine learning models in the drug discovery pipeline require large volumes of data for training and validation, particularly deep learning models.<sup>72</sup> The lack of adequate quality and robust data-sharing practices remain critical barriers for machine learning models to positively impact drug discovery.<sup>73</sup> Inadequate data quality can lead to models that have poor generalizability. Data harmonization, which improves the data quality and utilization via domain knowledge and machine learning techniques, plays a crucial role in the development and application of drug discovery.<sup>74</sup> Here, we briefly discuss several challenges and potential future directions as follows.

#### Interpretable generative models

While generative models and other deep learning-based approaches offer great potential, they are often essentially “black boxes” that require objective algorithmic interpretation of the predictions to provide confidence and actionability. Drug discovery is a highly complex process involving interactions between compounds and targets and interconnected biological systems. Current deep generative models are limited to capturing shallow statistical correlations of the data, which cannot explain mechanisms and results, possibly misleading decisions. Thus, model users must understand how the algorithms are constructed, which data they rely on, and to what extent the models are reliable. It is also important for AI scientists to involve biologists and clinicians in experimental design and data interpretation.

Models should be made interpretable.<sup>75</sup> One way is to perturb the input or parameters in the model and observe how the results change. For example, controllable molecular generation can be achieved by disentanglement, which decomposes the latent space into interpretable and independent factors that correspond to each property,<sup>76</sup> such as bioactivity and synthesizability. In this way, molecules with desired properties can be generated. Another solution can be displaying more semantic information from the algorithm to explain the causality of the results. The reasoning of relationships between molecular

structures and drug-like properties may guide the construction of causal graphs followed by molecule generation. Models can also be made transparent. Algorithms rationalize their prediction processes in a way that a human can understand. A hierarchical generative model may better trace each step back to previous levels, allowing for human-computer interaction to achieve targeted optimization.<sup>77</sup>

### Few-shot generative models

Current AI techniques rely on learning from large amounts of data. However, the available data are often quantitatively imbalanced due to, e.g., privacy, security, ethics,<sup>78</sup> or a small number of patients suffering from rare diseases, leading to little clinical data about the toxicity and poor bioactivity. Such situations could be alleviated by machines that learn from few samples. Combined with past knowledge, they can achieve good performance. Here, we highlight strategies to address insufficient data.

Starting from the source is the intuitive way to solve problems. Increasing the sample size can be achieved through data augmentation. Some approaches change the starting atom and the branching order in SMILES to enrich the data, taking advantage of the non-uniqueness of SMILES sequences for a structure.<sup>79</sup> Graph-based data can be varied by adding or removing edges using appropriate strategies,<sup>80</sup> such as 3D conformations.<sup>81</sup> This can be compounded by information at different granularity (e.g., atomic, pharmacophore, and toxicophore levels).

Insufficient training data of specific targets is inevitable in *de novo* molecular generation, especially for peptide or protein design. Transfer learning aims to transfer knowledge learned from one domain to a target domain related to the source domain, as solving data scarcity of the target domain.<sup>82</sup> Transfer learning drives molecule generation toward desired properties commonly in a fine-tuning manner from a pre-trained model.<sup>83</sup> The parameters obtained from the pre-trained model serve as the initialization of the specific task.

If no bioactive molecules are available, zero-shot learning, where a model can learn to recognize effects, or conditions, that were not observed, can be employed. Zero-shot learning requires more knowledge and alleviates the dependence on data. In rare diseases or orphan targets, learning compound-target interactions from big datasets, such as ChEMBL,<sup>12</sup> and designing molecules through disease-related targets instead of fitting molecular distributions, builds on “understanding the drug-target interactions.”

Considering that AlphaFold has uncovered 98.5% of human protein structures,<sup>84</sup> the target-based molecule generation can be converted into a classical image captioning problem. For example, image is the distance map (or 3D image) for a protein and captioning is the molecular SMILES code to be generated. In this configuration, target-based molecule generation can generally be handled with pipelines composed of a target visual encoder and a language model for SMILES generation.

### Multimodal generative models

The promise of successful drug discovery lies in the diversity of multiple data modalities that offer complementary perspectives and enable triangulating the evidence for discovery.<sup>85</sup> Deep

generative models using multimodal data may have significant advantages over unimodal counterparts since the multimodal data contain complementary insights.<sup>77</sup> Current studies usually focus on the molecular structural data, and do not fully use other data modalities, such as drug-target interactions, drug-disease knowledge, and relevant gene expression in specific cells following drug treatment (Figure 4A). Therefore, how to make full use of diverse and heterogeneous biological data is a matter worth discussing. There are multiple possible solutions to this challenge. First is “modality alignment,” which means connecting all modalities with an intermediate modality. Because establishing relationships with molecular structures is easier, the structure modality is chosen as the intermediary to other modalities, such as drug-induced gene expression. We then connect the structure modality with other modalities and finally align all modalities in the middle space. “Modality fusion,” which drops the median modality converter, is another possibility. All modalities are directly mapped to a common latent space and indicated by a hybrid representation (Figure 4A). Different modalities describing the same molecules should be closer in the modality-shared space, while the same modalities reflecting diverse molecules should be farther apart.

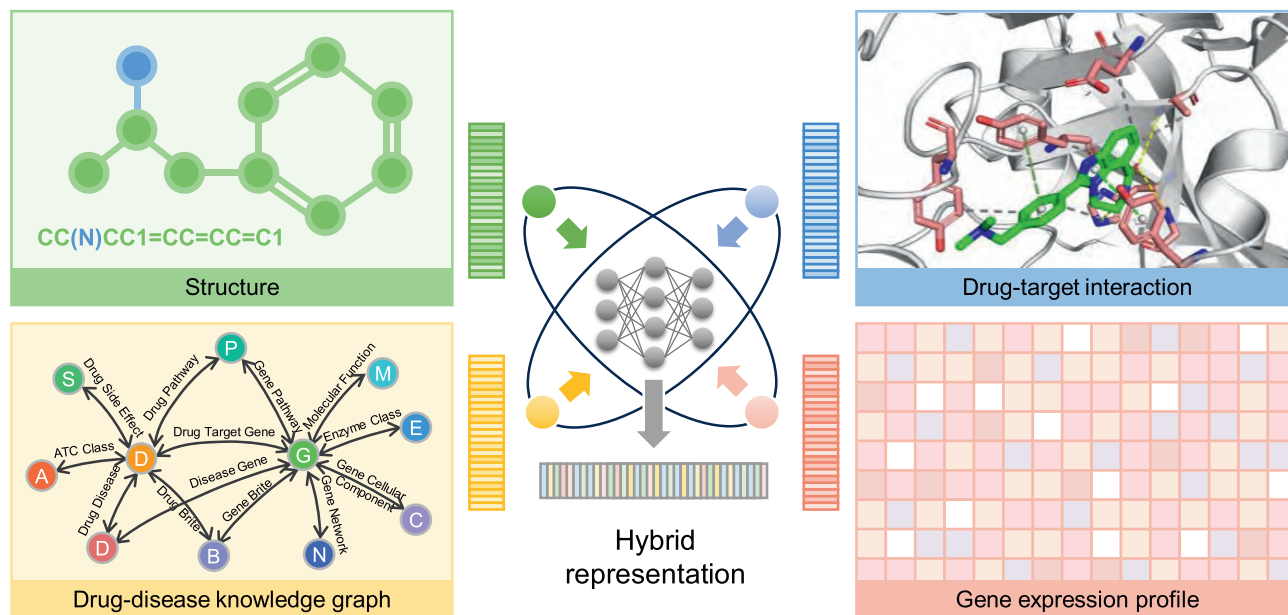
The above discussion is based on training data with sufficient and complete modalities, but the reality often does not satisfy such assumptions. To further exploit these partial data, we need to consider how to complement the missing modality. One possible way is to generate synthetic modalities through established relationships between modalities covering biological activities and pharmacokinetics and pharmacodynamics properties of molecules (Figure 4B). There is an urgent need to seek ways to integrate multimodal information that can generate molecules meaningfully to speed up the process of drug discovery.

### Generative models from data consumer to data producer

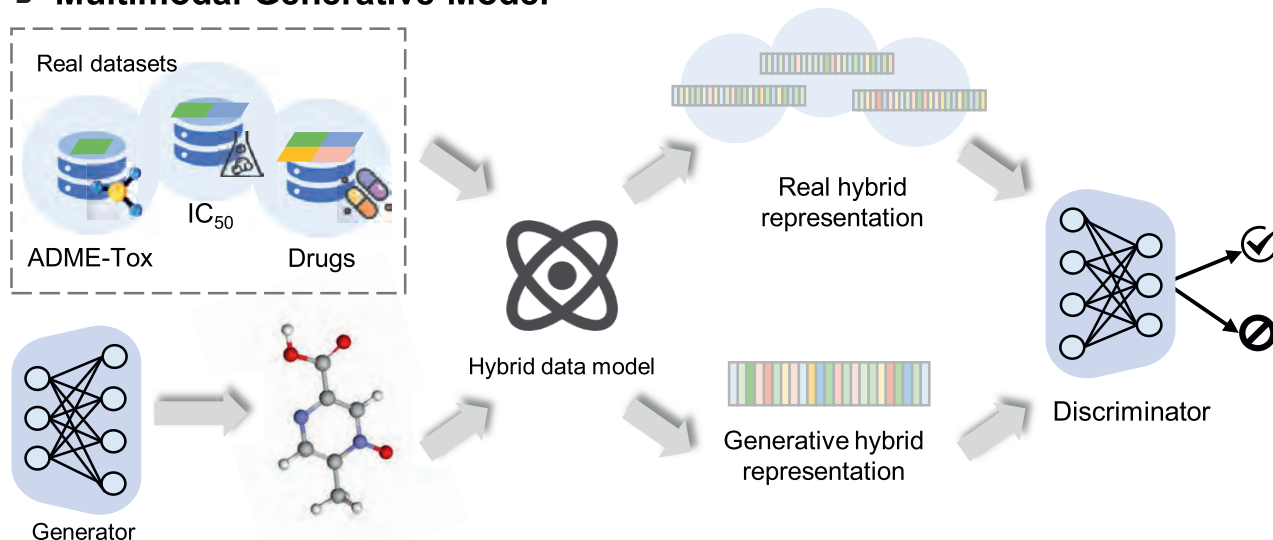
Unprecedented provision of data is pivotal to boosting data-driven drug discovery, in addition to the emergence of deep-learning algorithms and advances in high-performance computations based on the graphics processing unit. Pharmaceutical companies possess vast amounts of labeled data associated with their ~2–3M proprietary molecules and generated from the assays routinely run to support lead optimization. In addition, unlabeled data can be used for training as can computationally generated data such as from docking or molecular dynamics trajectories.<sup>86</sup>

The quantity of high-quality data<sup>87</sup> alone does not guarantee actionable decisions in drug discovery.<sup>88</sup> For example, leveraging a deep learning algorithm, AlphaFold predicts the 3D structure of proteins from their amino acid sequences and multi-sequence alignments with superior performance.<sup>30</sup> Yet critical details of the sites of molecular recognition, the active site for ligand binding or quaternary structure for protein-protein interaction, both vital for structure-based therapeutics design, remain unresolved. The affinity of the drug to the protein versus that of the substrate (or cofactor) determines its effectiveness. Yet, thermodynamic and dynamic properties are even farther from being routinely deployed in deep-learning models for drug design, despite their recognized importance. Free energy

## A Hybrid Data Model



## B Multimodal Generative Model



**Figure 4. A proposed multimodal generative model in the drug discovery applications**

(A) A hybrid data model can fully capture diverse information during drug design, including chemical, drug-target interactions, drug-disease knowledge, and disease-relevant expression of target (protein/gene).

(B) A multimodal generative model can consider various drug discovery pipeline components to increase likelihood of success of clinical trials. ADME-Tox, absorption, distribution, metabolism, and excretion-toxicity;  $IC_{50}$ , half-maximal inhibitory concentration.

calculations are frequently applied in lead optimization with a manageable size ( $> \sim 100$  s) of molecules, and, recently, protein-ligand binding kinetics have attracted attention in medicinal chemistry. However, the protein-ligand binding/unbinding dynamics is impractical to observe even in a long trajectory ( $\sim$ ms) from conventional molecular dynamics due to transition states separated by high energy barriers, thus locking the sys-

tem in configuration around its initial state, lacking conformational sampling.<sup>89</sup>

In this regard, a considerable effort employing deep-learning methods has been focused on enhanced samplings for extracting the free energy surface and kinetics, computing thermodynamics variables, constructing coarse-grained models, and generative modeling for molecular structure sampling.<sup>90</sup> For

example, a VAE-based generative network was employed to learn low-dimensional, non-linear embeddings by reconstructing time-lagged conformations, revealing the slow dynamics from the stochastic protein motions.<sup>91</sup> With a modified VAE in another example, weighted reaction coordinates optimized by maximizing a predictive information bottleneck framework can efficiently guide a biased simulation for capturing rare events in a short trajectory as well as calculating free energy and kinetics.<sup>92</sup>

Generative networks combined with molecular simulations solidly rooted in physics, could provide not only meaningful insights but also an invaluable framework for producing statistically reliable protein dynamics data for drug discovery, including COVID-19.<sup>93</sup> Still, in its infancy, it poses open questions, including some related to applications of generative modeling, e.g., accurate and efficient force field parameterization, enhanced sampling for kinetic modeling, and scalable generative modeling for a biological system. While current drug discovery is primarily devoted to small-molecule systems due to the data of proteins is severely limited, once the protein conformational dynamics data become more feasible, drug design would be driven toward enhanced safety and effectivity.

### Conclusions and outlook

Drug discovery platforms are becoming increasingly industrialized with the ability to both consume and generate big data using AI to drive new molecule design.<sup>94</sup> Ageing,<sup>95,96</sup> Alzheimer's disease,<sup>97,98</sup> COVID-19,<sup>6,65,93</sup> antimicrobial resistance,<sup>9</sup> and developments assisting the diagnosis and therapeutics of the COVID-19 pandemic<sup>6,99–101</sup> provide examples. These successes encourage us to embrace the challenges in further optimization and validation of AI approaches in medical applications. Increased enterprise architecture and infrastructure, including exascale computing,<sup>102</sup> quantum computers,<sup>103,104</sup> hardware, and connectivity, are a priority in drug discovery data strategies in industries, academia, and governments. Strong data stewardship practices enable the realization of interoperability and adherence to standards. Three rules have been highly recommended:

1. Data stewardship must ensure that data ownership rights (which lays the groundwork for data-sharing models) are operationalized and considered for data acquisition, use, and distribution practices.
2. Representative data (including diverse chemical and target coverage) is critical to ensuring the absence of data biases to allow deep learning models to cover a wide range of applications.
3. Big data's volume, variety, velocity, and veracity (4Vs) require automated and rigorous data harmonization and validation.

Data harmonization and validation from diverse biological endpoints and different assays can ensure data quality (completeness, consistency, integrity, fairness, and transparency) and data accuracy. In addition, advanced data-sharing and model-learning strategies, such as swarm learning<sup>105,106</sup>

and federated learning,<sup>74,107,108</sup> will accelerate data sharing among industries, academics, governments, and health care systems for drug development. For example, a recent platform called collaborative Profile-QSAR<sup>74</sup> developed collaborative models from previously reported biological assays to broaden the domain of applicability without sharing any of the training data, offering a way to address data scarcity.

In summary, recent advances triggered by the rapidly growing deep generative molecular design have brought new momentum for drug discovery, including the production and optimization of small molecules and macromolecules. However, the bottlenecks of AI technologies, such as lack of or limited interpretability of the model, inaccessibility, and lack of availability of high-quality data, currently restrict their application and affect their performance. There is a critical need to further develop and evaluate intelligent generative models in realistic real-world drug discovery contexts in order for deep learning to reach its full potential. Under such developments, the intelligent generative model paradigms will have the potential to transform from theoretical research to practical generation of therapeutics and provide easy-to-use toolkits for chemists and chemistry modelers in their daily work.

### ACKNOWLEDGMENTS

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract number HHSN261201500003I (to R.N.). This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, and Center for Cancer Research to R.N. This project was supported by the IBM-Cleveland Clinic Accelerator Initiative to F.C. and W.C.

### AUTHOR CONTRIBUTIONS

F.C. conceived the manuscript. X.Z., F.C., F.W., J.T., F.C.L., S.K., W.C., and E.F.F. contributed to critical discussion. X.Z. drafted the manuscript. X.Z., F.C., Y.L., S.K., W.C., and R.N. critically revised the manuscript.

### DECLARATION OF INTERESTS

E.F.F. has a CRADA arrangement with ChromaDex (USA) and is consultant to Aladdin Healthcare Technologies (UK and Germany), the Vancouver Dementia Prevention Centre (Canada), Intellectual Labs (Norway), and MindRank AI (China). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. S.K. and W.C. are employees of IBM TJ Watson Research Center. The other authors declare no competing interests.

### REFERENCES

1. Avorn, J. (2015). The \$2.6 billion pill—methodologic and policy considerations. *N. Engl. J. Med.* 372, 1877–1879.
2. Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature* 557, S55–S57.
3. Schütt, K.T., Gastegger, M., Tkatchenko, A., Müller, K.R., and Maurer, R.J. (2019). Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* 10, 5024.
4. Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., Fang, J., Huang, Y., Guo, H., Li, L., et al. (2020). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797.

5. Hie, B., Zhong, E.D., Berger, B., and Bryson, B. (2021). Learning the language of viral evolution and escape. *Science* 371, 284–288.
6. Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020). Artificial intelligence in COVID-19 drug repurposing. *Lancet. Digit. Health* 2, e667–e676.
7. Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Jr., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., et al. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19, 353–364.
8. Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* 15, 816–822.
9. Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell* 181, 475–483.
10. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276.
11. Irwin, J.J., Tang, K.G., Young, J., Dandarchuluun, C., Wong, B.R., Khur-elbaatar, M., Moroz, Y.S., Mayfield, J., and Sayle, R.A. (2020). ZINC20-A free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* 60, 6065–6073.
12. Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J.P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.
13. Ruddigkeit, L., van Deursen, R., Blum, L.C., and Reymond, J.L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875.
14. Patel, H., Ihlenfeldt, W.D., Judson, P.N., Moroz, Y.S., Pevzner, Y., Peach, M.L., Delannée, V., Tarasova, N.I., and Nicklaus, M.C. (2020). SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* 7, 384.
15. Hoffmann, T., and Gastreich, M. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov. Today* 24, 1148–1156.
16. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein Data Bank. *Nucleic Acids Res.* 28, 235–242.
17. Weininger, D. (1988). A chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* 28, 31–36.
18. Schwalbe-Koda, D., and Gómez-Bombarelli, R. (2020). Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics* (Springer), pp. 445–467.
19. Gupta, N., Mangal, N., and Biswas, S. (2005). Evolution and similarity evaluation of protein structures in contact map space. *Proteins* 59, 196–204.
20. David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* 12, 56.
21. Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M.M., and Correia, B.E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* 17, 184–192.
22. Wójcikowski, M., Kukielka, M., Stepniewska-Dziubinska, M.M., and Siedlecki, P. (2019). Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* 35, 1334–1341.
23. Mahmoud, A.H., Masters, M.R., Yang, Y., and Lill, M.A. (2020). Elucidating the multiple roles of hydration for accurate protein–ligand binding prediction via deep learning. *Commun. Chem.* 3, 19.
24. Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W.F.D., Kirshner, D., Wong, S.E., Lightstone, F.C., and Allen, J.E. (2021). Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* 61, 1583–1592.
25. Xu, M., Wang, W., Luo, S., Shi, C., Bengio, Y., Gomez-Bombarelli, R., and Tang, J. (2021). An end-to-end framework for molecular conformation generation via bilevel programming. In *International Conference on Machine Learning (PMLR)*, pp. 11537–11547.
26. Shi, C., Luo, S., Xu, M., and Tang, J. (2021). Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning (PMLR)*, pp. 9558–9568.
27. Axelrod, S., and Gómez-Bombarelli, R. (2022). GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* 9, 185–214.
28. Imrie, F., Hadfield, T.E., Bradley, A.R., and Deane, C.M. (2021). Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* 12, 14577–14589.
29. Li, Y., Pei, J., and Lai, L. (2021). Structure-based de novo drug design using 3D deep generative models. *Chem. Sci.* 12, 13664–13675.
30. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589.
31. Sun, Z., Zhu, Q., Mou, L., Xiong, Y., Li, G., and Zhang, L. (2019). A grammar-based structural cnn decoder for code generation. *Proc. AAAI Conf. Artif. Intell.* 33, 7055–7062.
32. Hadjeres, G., and Nielsen, F. (2020). Enforcing unary constraints in sequence generation, with application to interactive music generation. *Neural Comput. Appl.* 32, 995–1005.
33. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780.
34. Cho, K., Merriënboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar (ACL)*. A meeting of SIGDAT, a special interest Group of the ACL 1724–1734.
35. Brown, N., Fiscato, M., Segler, M.H.S., and Vaucher, A.C. (2019). Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* 59, 1096–1108.
36. Mita, G., Filippone, M., and Michiardi, P. (2021). An identifiable double VAE for disentangled representations. In *International Conference on Machine Learning (PMLR)*, pp. 7769–7779.
37. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144.
38. Rezende, D., and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning (PMLR)*, pp. 1530–1538.
39. Zang, C., and Wang, F. (2020). MoFlow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 617–626.
40. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature* 550, 354–359.
41. Popova, M., Isayev, O., and Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4, eaap7885.

42. Ertl, P., and Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* *1*, 8.
43. Wang, J., Xu, P., Hao, Y., Yu, T., Liu, L., Song, Y., and Li, Y. (2021). Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *BMC Cancer* *21*, 914–922.
44. Gottipati, S.K., Sattarow, B., Niu, S., Pathak, Y., Wei, H., Liu, S., Thomas, K.M.J., Blackburn, S., Coley, C.W., Tang, J., et al. (2020). Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning (PMLR)*, pp. 3668–3679.
45. Kitchen, D.B., Decornez, H., Furr, J.R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* *3*, 935–949.
46. Bleicher, K.H., Böhm, H.J., Müller, K., and Alanine, A.I. (2003). Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* *2*, 369–378.
47. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* *23*, 1241–1250.
48. Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. (2018). Syntax-directed variational autoencoder for molecule generation. In *Proceedings of the International Conference on Learning Representations*.
49. Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning (PMLR)*, pp. 2323–2332.
50. Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. (2018). Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR)*.
51. Jacobs, S.A., Moon, T., McLoughlin, K., Jones, D., Hysom, D., Ahn, D.H., Gyllenhaal, J., Watson, P., Lightstone, F.C., Allen, J.E., et al. (2021). Enabling rapid COVID-19 small molecule drug design through scalable deep learning of generative models. *Int. J. High Perform. Comput. Appl.* *35*, 469–482.
52. Kuznetsov, M., and Polykovskiy, D. (2021). MolGrow: a graph normalizing flow for hierarchical molecular generation. *Proc. AAAI Conf. Artif. Intell.* *35*, 8226–8234.
53. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* *11*, 1–10.
54. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* *37*, 1038–1040.
55. Jin, W., Barzilay, R., and Jaakkola, T. (2020). Multi-objective molecule generation using interpretable substructures. In *International Conference on Machine Learning (PMLR)*, pp. 4849–4859.
56. Beker, W., Wołos, A., Szymkuć, S., and Grzybowski, B.A. (2020). Minimal-uncertainty prediction of general drug-likeness based on Bayesian neural networks. *Nat. Mach. Intell.* *2*, 457–465.
57. Jin, W., Yang, K., Barzilay, R., and Jaakkola, T.S. (2019). Learning multi-modal graph-to-graph translation for molecule optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019 (OpenReview.net)*.
58. Zhu, J.-Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.
59. Maziarka, Ł., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., and Warchol, M. (2020). Mol-CycleGAN: a generative model for molecular optimization. *J. Cheminform.* *12*, 2–18.
60. Cadow, J., Born, J., Manica, M., Oskooei, A., and Rodríguez Martínez, M. (2020). A web service for interpretable anticancer compound sensitivity prediction. *Nucleic Acids Res.* *48*, W502–W508.
61. Born, J., Huynh, T., Stroobants, A., Cornell, W.D., and Manica, M. (2021). Active site sequence representations of human kinases outperform full sequence representations for affinity prediction and inhibitor generation: 3D effects in a 1D model. *J. Chem. Inf. Model.* *62*, 240–257.
62. Ghosh, D., Veeraraghavan, B., Elangovan, R., and Vivekanandan, P. (2020). Antibiotic resistance and epigenetics: more to it than meets the eye. *Antimicrob. Agents Chemother.* *64*. 022255-e19.
63. Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning (PMLR)*, pp. 214–223.
64. Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Dos Santos, C., Chen, P.Y., et al. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* *5*, 613–623.
65. Linsky, T.W., Vergara, R., Codina, N., Nelson, J.W., Walker, M.J., Su, W., Barnes, C.O., Hsiang, T.Y., Esser-Nobis, K., Yu, K., et al. (2020). De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science* *370*, 1208–1214.
66. Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynešas, A., Viknander, S., Abuajwa, W., et al. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* *3*, 324–333.
67. Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B., et al. (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* *19*, 80.
68. Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., Lorenzin, F., Prandi, D., Romanel, A., Demichelis, F., et al. (2018). A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* *36*, 265–271.
69. Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., and Wang, Y. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* *10*, 4284–4314.
70. Gelijns, A.C. (1989). Institute of Medicine Committee on Technological Innovation in. *M. Technological Innovation: Comparing Development of Drugs, Devices, and Procedures in Medicine (National Academies Press)*.
71. Austin, C.P. (2021). Opportunities and challenges in translational science. *Clin. Transl. Sci.* *14*, 1629–1647.
72. AlQuraishi, M., and Sorger, P.K. (2021). Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. Methods* *18*, 1169–1180.
73. Bender, A., and Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* *26*, 1040–1052.
74. Martin, E.J., and Zhu, X.W. (2021). Collaborative profile-QSAR: a natural platform for building collaborative models among competing companies. *J. Chem. Inf. Model.* *61*, 1603–1616.
75. Weber, J.K., Morrone, J.A., Bagchi, S., Pabon, J.D.E., Kang, S.G., Zhang, L., and Cornell, W.D. (2022). Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. *J. Comput. Aided Mol. Des.* *36*, 391–404.
76. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). Beta-VAE: learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings (OpenReview.net)*.
77. Manica, M., Oskooei, A., Born, J., Subramanian, V., Sáez-Rodríguez, J., and Rodríguez Martínez, M. (2019). Toward explainable anticancer



- compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.* **16**, 4797–4806.
78. Wang, Y., Yao, Q., Kwok, J.T., and Ni, L.M. (2020). Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surv.* **53**, 1–34.
  79. Arús-Pous, J., Johansson, S.V., Prykhodko, O., Bjerrum, E.J., Tyrchan, C., Reymond, J.L., Chen, H., and Engkvist, O. (2019). Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminform.* **11**, 71.
  80. Zhao, T., Liu, Y., Neves, L., Woodford, O., Jiang, M., and Shah, N. (2021). Data augmentation for graph neural networks. *Proc. AAAI Conf. Artif. Intell.* **35**, 11015–11023.
  81. Hemmerich, J., Asilar, E., and Ecker, G.F. (2020). COVER: conformational oversampling as data augmentation for molecules. *J. Cheminform.* **12**, 18.
  82. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76.
  83. Segler, M.H.S., Kogej, T., Tyrchan, C., and Waller, M.P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131.
  84. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Židek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596.
  85. Luo, Y., Eran, A., Palmer, N., Avillach, P., Levy-Moonshine, A., Szolovits, P., and Kohane, I.S. (2020). A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nat. Med.* **26**, 1375–1379.
  86. Bayarri, G., Hospital, A., and Orozco, M. (2021). 3dRS, a web-based tool to share interactive representations of 3D biomolecular structures and molecular dynamics trajectories. *Front. Mol. Biosci.* **8**, 726232.
  87. Nigam, A., Pollice, R., Hurley, M.F.D., Hickman, R.J., Aldeghi, M., Yoshikawa, N., Chithrananda, S., Voelz, V.A., and Aspuru-Guzik, A. (2021). Assigning confidence to molecular property prediction. *Expert Opin. Drug Discov.* **16**, 1009–1023.
  88. Bender, A., and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov. Today* **26**, 511–524.
  89. Allison, J.R. (2020). Computational methods for exploring protein conformations. *Biochem. Soc. Trans.* **48**, 1707–1724.
  90. Noé, F., Tkatchenko, A., Müller, K.R., and Clementi, C. (2020). Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390.
  91. Wehmeyer, C., and Noé, F. (2018). Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **148**, 241703.
  92. Wang, Y., Ribeiro, J.M.L., and Tiwary, P. (2019). Past-future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* **10**, 3573.
  93. Sztain, T., Ahn, S.H., Boggetti, A.T., Casalino, L., Goldsmith, J.A., Seitz, E., McCool, R.S., Kearns, F.L., Acosta-Reyes, F., Maji, S., et al. (2021). A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat. Chem.* **13**, 963–968.
  94. Sadybekov, A.A., Sadybekov, A.V., Liu, Y., Iliopoulos-Tsoutsouvas, C., Huang, X.P., Pickett, J., Houser, B., Patel, N., Tran, N.K., Tong, F., et al. (2022). Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459.
  95. Aman, Y., Frank, J., Lautrup, S.H., Matysek, A., Niu, Z., Yang, G., Shi, L., Bergersen, L.H., Storm-Mathisen, J., Rasmussen, L.J., et al. (2020). The NAD(+)-mitophagy axis in healthy longevity and in artificial intelligence-based clinical applications. *Mech. Ageing Dev.* **185**, 111194.
  96. Mkrtchyan, G.V., Abdelmohsen, K., Andreux, P., Bagdonaite, I., Barzilai, N., Brunak, S., Cabreiro, F., de Cabo, R., Campisi, J., Cuervo, A.M., et al. (2020). Ardd 2020: from aging mechanisms to interventions. *Aging (Albany NY)* **12**, 24484–24503.
  97. Fang, J., Zhang, P., Zhou, Y., Chiang, C.W., Tan, J., Hou, Y., Stauffer, S., Li, L., Pieper, A.A., Cummings, J., and Cheng, F. (2021). Endophenotype-based in-silico network medicine discovery combined with insurance records data mining identifies sildenafil as a candidate drug for Alzheimer's disease. *Nat. Aging* **1**, 1175–1188.
  98. Taubes, A., Nova, P., Zalocusky, K.A., Kosti, I., Bicak, M., Zilberter, M.Y., Hao, Y., Yoon, S.Y., Oskotsky, t., Pineda, S., et al. (2021). Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nat. Aging* **1**, 932–947.
  99. Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., and Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* **6**, 14.
  100. Zhou, Y., Hou, Y., Shen, J., Mehra, R., Kallianpur, A., Culver, D.A., Gack, M.U., Farha, S., Zein, J., Comhair, S., et al. (2020). A network medicine approach to prediction and population-based validation of disease manifestations and drug repurposing for COVID-19. *PLoS Biol.* **18**, e3000970.
  101. Galindez, G., Matschinske, J., Rose, T.D., Sadegh, S., Salgado-Albarrán, M., Späth, J., Baumbach, J., and Pauling, J.K. (2021). Lessons from the COVID-19 pandemic for advancing computational drug repurposing strategies. *Nat. Comput. Sci.* **1**, 33–41.
  102. Nussinov, R., Jang, H., Nir, G., Tsai, C.J., and Cheng, F. (2021). A new precision medicine initiative at the dawn of exascale computing. *Signal Transduct. Target. Ther.* **6**, 3.
  103. Abbott, A. (2021). Quantum computers to explore precision oncology. *Nat. Biotechnol.* **39**, 1324–1325.
  104. Satzinger, K.J., Liu, Y.J., Smith, A., Knapp, C., Newman, M., Jones, C., Chen, Z., Quintana, C., Mi, X., Dunsworth, A., et al. (2021). Realizing topologically ordered states on a quantum processor. *Science* **374**, 1237–1241.
  105. Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., et al. (2021). Swarm Learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270.
  106. Ferrer, E.C., Hardjono, T., Pentland, A., and Dorigo, M. (2021). Secure and secret cooperation in robot swarms. *Sci. Robot.* **6**, eabf1538.
  107. Chen, S., Xue, D., Chuai, G., Yang, Q., and Liu, Q. (2021). A federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics* **36**, 5492–5498.
  108. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119.

## Review

# From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment

Kyle Swanson,<sup>1,6</sup> Eric Wu,<sup>2,6</sup> Angela Zhang,<sup>3,6</sup> Ash A. Alizadeh,<sup>4</sup> and James Zou<sup>1,2,5,\*</sup>

<sup>1</sup>Department of Computer Science, Stanford University, Stanford, CA, USA

<sup>2</sup>Department of Electrical Engineering, Stanford University, Stanford, CA, USA

<sup>3</sup>Department of Genetics, Stanford University, Stanford, CA, USA

<sup>4</sup>Department of Medicine, Stanford University, Stanford, CA, USA

<sup>5</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

<sup>6</sup>These authors contributed equally

\*Correspondence: jamesz@stanford.edu

<https://doi.org/10.1016/j.cell.2023.01.035>

## SUMMARY

Machine learning (ML) is increasingly used in clinical oncology to diagnose cancers, predict patient outcomes, and inform treatment planning. Here, we review recent applications of ML across the clinical oncology workflow. We review how these techniques are applied to medical imaging and to molecular data obtained from liquid and solid tumor biopsies for cancer diagnosis, prognosis, and treatment design. We discuss key considerations in developing ML for the distinct challenges posed by imaging and molecular data. Finally, we examine ML models approved for cancer-related patient usage by regulatory agencies and discuss approaches to improve the clinical usefulness of ML.

## INTRODUCTION

In the past decade, machine learning (ML) has seen an explosion of applications in medicine, particularly within oncology.<sup>1</sup> As a set of complex, heterogeneous, and prevalent diseases, cancers provide both a challenging set of diagnostic problems and copious data in multiple modalities.<sup>2</sup> This makes clinical oncology a promising field for ML, which utilizes data to learn patterns and the structure of a dataset (see machine learning primer section for a brief introduction to ML). In particular, rich imaging and molecular data have spurred the application of ML to correlate these data sources with early cancer detection, monitoring of cancer progression, and identification of optimized therapeutic treatment.

Medical imaging has been a powerful tool that has revolutionized cancer diagnostics. In particular, medical imaging enables non-invasive, cheap, and scalable detection, localization, and monitoring of cancer. Radiology images, as well as other image modalities like skin images or colonoscopy videos, are used for screening and diagnosis.<sup>3</sup> Pathology images of tissue samples are used to confirm a cancer diagnosis and determine prognostic factors such as cancer subtype.<sup>4</sup> Both radiology and pathology images can guide treatment by informing the selection of chemotherapy or immunotherapy and aiding radiotherapy planning.<sup>5</sup> As medical imaging is increasingly fundamental to the clinical oncology workflow, the quantity of imaging data is often growing faster than clinicians can handle.<sup>3</sup> This leads to a desire for automated methods of processing medical images to reduce clinician workload, accelerate the analysis of time-sensitive

images, and mitigate clinician errors. Advances in ML for computer vision have been adapted for medical imaging and are already showing great promise for rapidly and accurately analyzing a variety of imaging modalities in clinical oncology.<sup>6,7</sup>

Although imaging informs many aspects of cancer care, molecular characterization can provide a more fine-grained view of a patient's cancer status.<sup>8</sup> This is particularly important as cancer therapeutics become increasingly targeted and mechanistic.<sup>9</sup> Liquid biopsies, which measure molecular biomarkers present in non-invasive physiology samples such as blood or urine, have emerged as a promising approach to profiling tumor states for cancer diagnostics. Liquid and solid tumor biopsies also make it possible to serially profile tumor status and identify characteristics of tumor evolution and heterogeneity that are associated with resistance to therapies, recurrence, and poor survival outcomes.<sup>10</sup> Due to the wealth of information provided by liquid biopsies and solid tumor biopsies, ML has been instrumental in predicting clinical outcomes and cancer status from rich molecular features.

In this review, we explore recent advances in ML applied to clinical oncology. We focus on relatively mature ML technologies already deployed or close to deployment in clinical settings. There is a large body of exciting development of ML for more basic cancer research and drug discovery that we do not cover here. Because imaging and molecular data are two major data modalities in clinical oncology with distinct ML challenges, we structure the review to discuss imaging ML and molecular ML separately. For each modality, we discuss both the major applications of ML and the types of ML models and techniques that



are frequently used. As many of these ML models are moving from lab to clinic, we also review the regulatory process for approving ML methods for cancer diagnostics. We highlight examples of recently approved ML-based devices in this category and discuss the clinical studies necessary to obtain approval. We then discuss how to improve ML model design and evaluation in order to build trust in cancer-related ML and further clinical adoption. Finally, we outline emerging technologies, both in medicine and ML, that are promising directions for future research in clinical oncology.

## MACHINE LEARNING PRIMER

ML aims to solve tasks by learning patterns from data rather than using hand-coded rules.<sup>4</sup> An ML model is trained to perform a task by showing it several examples of input data (e.g., mammograms) and corresponding output labels (e.g., cancer or no cancer) and updating the internal parameters of the model accordingly to make its predictions more accurate. Model evaluation on external test data, which comes from an entirely different source than the training and internal test data (e.g., a different hospital or patient population), is particularly valuable to determine the model's generalizability across diverse settings. While most ML methods for cancer are a form of supervised learning, where each data point has an associated label, unsupervised learning methods such as clustering and dimensionality reduction can produce relevant insights into unlabeled data.<sup>7</sup>

### Traditional ML vs. deep learning

Traditional ML algorithms take a wide variety of forms, with most designed to work with tabular data, where each data point has a set of explicit features (e.g., patient age or gene mutation status) that are used to predict the label.<sup>3</sup> One common algorithm is called a random forest, which consists of a set of decision trees, each of which is constructed based on the training data to make a series of binary decisions about the input features that culminates in a prediction of the label of the data point. Another algorithm is the support vector machine (SVM), which learns a line (or hyperplane in multiple dimensions) in the coordinate system defined by the input features to separate the data points into two classes. Regression models learn a linear combination of input features that predict either continuous labels (e.g., linear regression) or binary labels (e.g., logistic regression).

With the increasing availability and power of graphics processing units (GPUs), a subfield of machine learning called deep learning (DL) has overtaken traditional ML for many prediction tasks.<sup>3</sup> The core component of DL models is a neural network, which consists of one or more layers of units called neurons that compute weighted sums of inputs followed by applying a nonlinear function. These layers of neurons thus compute a representation of the input called an embedding, which is then used by the final layer of neurons to make an output prediction. The DL models are more flexible compared to traditional ML models, and because DL relies less on feature engineering, they are capable of processing a wider variety of unstructured data types including images, text, and speech. However, DL models typically require significantly more training

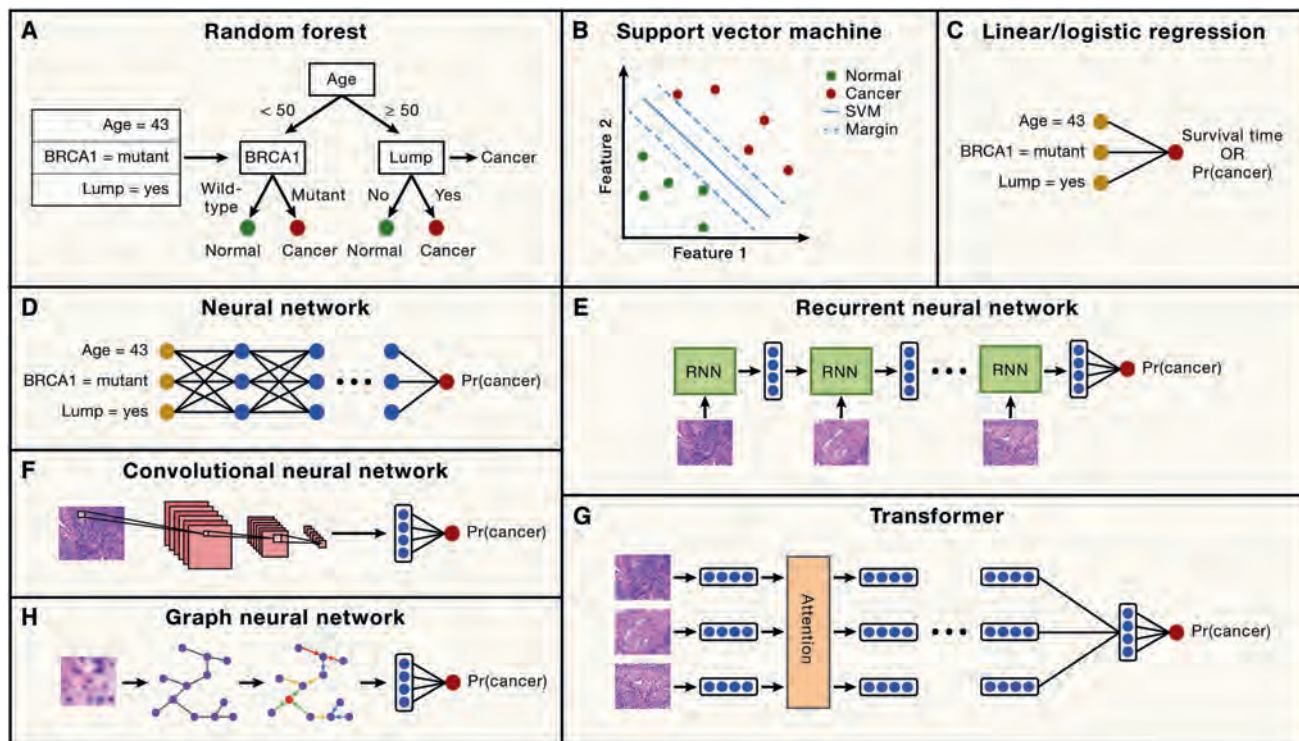
data, so traditional ML models can still be useful, particularly for data-limited or tabular tasks.<sup>2</sup>

In order to process non-tabular data, the architecture of a neural network (e.g., number of neurons or layers or connections between neurons) is modified to fit the desired data type.<sup>2</sup> Convolutional neural networks (CNNs) are primarily designed for processing images. Graph neural networks (GNNs) handle graph data, such as cell-cell interaction graphs. Recurrent neural networks (RNNs) and transformers analyze sequential data, such as genetic sequences or series of images. Each of these classes of models has many specific model architectures, such as ResNet or U-Net for CNNs and LSTM or GRU for RNNs. The models are optimized with stochastic gradient descent. Figure 1 illustrates common traditional ML and DL models.

Both traditional ML and DL models require that the data is cleaned (e.g., modifying data with missing features or extreme values) in order to learn effectively.<sup>4</sup> Additionally, the input features must be amenable to the type of model. For example, neural networks use vectors of real numbers as input, so categorical features such as cancer type are typically converted to one-hot vectors with all zeros except for a single one in a position that indicates the appropriate category. Many traditional ML methods are available in the scikit-learn package, while deep learning models can be built using packages such as PyTorch and TensorFlow. Because ML models often require tuning hyperparameters to obtain optimal performance, it is important that a validation dataset be used during this step that is distinct from the held-out test dataset, which is only evaluated after the final hyperparameters have been chosen.

### Training techniques

One common technique is transfer learning, where a model is first trained on a large dataset that is somewhat related to the task of interest (pre-training) before being trained on a smaller dataset consisting of the actual task of interest (fine-tuning).<sup>3</sup> For example, image-based cancer detection models are often pre-trained on large object detection datasets, enabling the model to recognize general shapes, and are then fine-tuned on small cancer detection image datasets. Transfer learning is more useful when the pre-training data are similar to the data of interest. Another common method is data augmentation, where input data are modified (e.g., images are rotated or blurred) to artificially expand the training set and make the model more robust to noise that might appear in real-world data.<sup>7</sup> Regularization is a technique that controls the size of the parameters of a model to prevent overfitting and encourage sparse feature usage.<sup>2</sup> Weak supervision involves using data with limited or noisy label information.<sup>7</sup> A common type of weak supervision is multiple instance learning, in which labeled data points (e.g., images with cancer/no cancer labels) are broken down into smaller pieces (e.g., image tiles) that are easier for an ML model to process. The model makes predictions on each piece of the data separately, and those predictions are then aggregated to form a prediction for the whole data point. Finally, interpretability is a set of methods that aim to explain why a model is making a certain prediction.<sup>6</sup> For example, an image-based model might highlight regions of an image that led the model to diagnose a patient with cancer.



**Figure 1. Common machine learning models**

(A) A random forest model builds decision trees that make predictions based on a series of binary decisions about the input features.

(B) A support vector machine (SVM) learns a line (or hyperplane in many dimensions) in feature space that separates two classes of data points with the largest possible margin between the two classes.

(C) A regression model uses a linear combination of input features to predict either continuous labels (linear regression) or binary labels (logistic regression).

(D) A neural network consists of multiple layers of neurons that iteratively compute linear combinations of inputs followed by a nonlinear function to predict outcomes such as the probability of cancer.

(E) An RNN processes sequential data, such as genetic sequences or a series of images, by applying the same neural network layers to each object in the sequence and maintaining a memory of the objects it has seen.

(F) A CNN applies patches of neurons called filters that scan an image for patterns. Early layers identify low-level features like edges, while later layers identify high-level features such as tumor morphology.

(G) A transformer analyzes sequential data by repeatedly applying an operation called attention to compare each element in the sequence to all the other elements in order to update its internal representation of the sequence.

(H) A GNN is designed for graph-structured data such as a graph of neighboring cells. It first encodes basic features of each node and edge in the graph, and then neural network layers pass information across the graph to update the node and edge representations, which are then used to predict the label of the graph. Each of these general classes of models has many specific architectures with different numbers and sizes of layers of neurons.

Image sources: histology.<sup>11</sup>

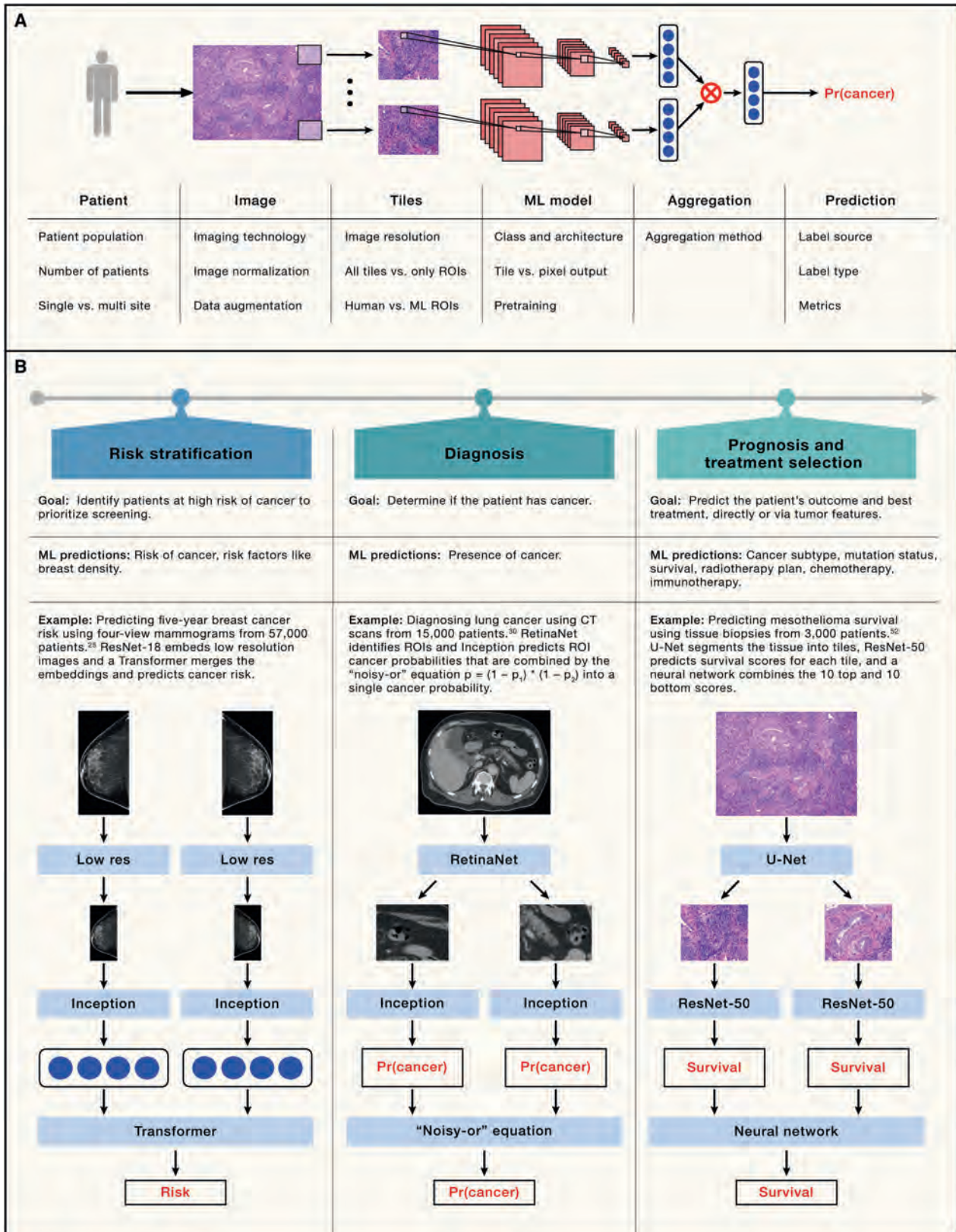
## MACHINE LEARNING FOR IMAGE-BASED CANCER DIAGNOSIS, PROGNOSIS, AND TREATMENT

In this section, we highlight applications of image-based ML throughout the clinical workflow for cancer. Early ML approaches used hand-crafted image features such as tumor shape or textural heterogeneity that were computationally extracted from images.<sup>6</sup> These features were used as inputs to a traditional ML model, such as a support vector machine (SVM) or random forest, to make a clinical prediction. Starting in the early 2010s, a class of ML models called deep learning (DL) models began to take hold as the dominant ML method.<sup>12</sup> DL models automatically learn features from an image to make clinical predictions, thereby simultaneously reducing the need for painstakingly crafting image features while significantly outperforming feature-based ML models.<sup>3,4</sup> These models can be

applied to virtually any medical imaging modality, including X-ray<sup>13</sup> and MRI for radiology,<sup>14</sup> H&E stains for pathology,<sup>15</sup> images of skin lesions for dermatology,<sup>16</sup> and videos of colonoscopies for gastroenterology.<sup>17</sup> Here, we discuss examples of ML—primarily DL—applied to three clinical stages: risk stratification, diagnosis, and prognosis and treatment planning. Figure 2 illustrates the general image-based ML model pipeline and each of the three clinical stages. Although we discuss each stage separately, it is worth noting that some ML methods make predictions that cross these boundaries, such as simultaneous diagnosis and prognosis via pathology images.<sup>18</sup>

### Risk stratification

Understanding a patient's risk of developing cancer is important for early cancer detection and effective treatment. Often, cancer risk is evaluated based on a patient's demographics, family



(legend on next page)

history, and genetics, but imaging can also reveal patient characteristics that might increase cancer risk. Existing work on image-based cancer risk prediction falls into two categories: predicting characteristics associated with cancer risk and directly predicting cancer risk itself.

### Risk proxies

A typical example of a characteristic associated with cancer risk is breast density in breast cancer. Breast density is correlated with increased risk of cancer development and missed detection on mammography and therefore indicates who may benefit from additional screening.<sup>21</sup> To improve breast density assessment with DL, Lehman et al. trained a ResNet-18 CNN model on mammograms to predict breast density categories routinely evaluated in clinical practice.<sup>21</sup> The model showed a high level of agreement with a panel of five radiologists on a held-out test set of images. Furthermore, the DL model was implemented in clinical practice, and radiologists accepted the binary density predictions of the model 94% of the time. The model was additionally validated at an external site and showed the potential to increase the consistency of breast density evaluations by radiologists at different sites.<sup>22</sup>

### Risk prediction

More often than quantifying risk proxies, DL is used to directly predict cancer risk. For example, DL models are often trained to use images from a screening mammogram to predict whether a patient will develop cancer at some point.<sup>23</sup> Dembrower et al. highlight the benefit of this direct approach to risk prediction, as they showed that a breast cancer risk score produced by an Inception-ResNet-v2 CNN model was more accurate than using clinical breast density assessments to predict risk.<sup>24</sup> Yala et al. developed a DL model on mammograms that could better predict the likelihood that a woman would develop breast cancer within five years than the well-established Tyrer-Cuzick risk model, which is based on clinical features such as patient age.<sup>25</sup> Their method consisted of a ResNet-18 model to process each of the four standard mammogram views, followed by a transformer network that aggregated the view embeddings into a single mammogram embedding. This embedding was used to predict known risk factors, a baseline cancer risk score, and a hazard score for additional risk in future years. They also used a conditional-adversarial training scheme to make the model invariant to the mammogram device to ensure consistent risk assessments across devices. The authors later validated their model on test sets from seven hospitals across five countries, demonstrating the generalizability of the model across

diverse patient populations and screening centers.<sup>26</sup> Ha et al. designed a CNN model that predicts risk not only at the image level but also at the pixel level, meaning that each risk prediction score comes with a heatmap on the image indicating the regions where cancer is most likely to develop.<sup>27</sup> Although most studies have focused on risk stratification for breast cancer, ML has also been used for predicting lung cancer risk from chest X-rays with CNNs<sup>13</sup> and for predicting prostate cancer risk from MRIs, with support vector machines applied to hand-crafted radiomics features.<sup>14</sup>

These methods aim to personalize cancer screening by providing a risk score to a physician, who is then responsible for determining an appropriate screening frequency for the patient. However, since standard, non-ML risk scores are relatively coarse-grained and imprecise, current guidelines place patients in large groups based on high or low risk and suggest the same screening schedule for all patients in a group, rather than adapting the screening frequency uniquely for each patient.<sup>28</sup> Yala et al. demonstrated that reinforcement learning, an area of machine learning that involves deciding which actions to take to maximize a reward, can be used in conjunction with DL risk prediction models to automatically design an optimal screening schedule for each patient individually.<sup>28</sup> These individual screening schedules significantly improved simulated early detection rates per screening mammogram compared to standard clinical guidelines.

### Diagnosis

Diagnosing cancer typically involves two steps. First, either in the course of routine screening or in response to symptoms, patients undergo non-invasive imaging such as radiological scans. Second, if these images reveal suspicious regions of tissue that might indicate cancer, a biopsy is then taken and sent to a pathology lab, which can confirm the diagnosis with the help of histological imaging. ML can improve the diagnostic accuracy of both of these steps by identifying patterns—both known and unknown to clinicians—that indicate the presence or absence of cancer. ML also provides a consistent and detailed image evaluation that can catch cancers missed by time-constrained physicians, which is particularly crucial in radiology for early detection.

### Non-invasive imaging

Detecting signs of cancer via ML applied to radiological or other non-invasive imaging has garnered substantial attention and excitement due to the abundance of data and the success of

## Figure 2. Machine learning for image-based cancer diagnosis, prognosis, and treatment

(A) An illustration of the general ML model pipeline for image-based cancer prediction tasks, along with key considerations at each step. For each patient in a patient population, an image is captured from radiology, pathology, or another imaging modality. Often, the image is high resolution and is broken down into image tiles—either covering the full image or only ROIs—that are small enough for an ML model to process. An ML model processes each image tile, producing an embedding of the tile or a tile-level or pixel-level prediction. The tile outputs are aggregated into a single output using either a formula or an ML model such as an RNN. A final prediction component, such as a neural network, uses the combined tile output to predict the label, and metrics evaluate the model predictions. Labels may come from different sources (e.g., radiology or biopsy) and can have different types (e.g., binary for classification or real-valued for regression). (B) The clinical stages of image-based ML predictions for cancer and simplified examples of ML methods for each stage.

**Risk Stratification:** For certain cancers such as breast cancer, healthy patients regularly undergo radiological screening to assess the patient's risk of developing cancer and prioritize future screening.

**Diagnosis:** Radiology images are used to identify potentially cancerous lesions during routine screening or in response to symptoms. If cancer is suspected by radiology, then a biopsy is taken, and pathology images are used to confirm the diagnosis.

**Prognosis and Treatment Selection:** Radiology or pathology images are further used to evaluate prognosis and select treatments.

Image sources: mammography,<sup>19</sup> CT,<sup>20</sup> histology.<sup>11</sup>

ML methods, with several claiming to achieve physician-level performance for cancer detection. These methods hold promise to improve and standardize early detection of cancer, save physicians time, and expand access to high-quality cancer care to patients in low-resource settings. Esteva et al. trained an Inception v3 CNN to classify skin cancer from images of skin lesions, matching the performance of 21 dermatologists on biopsy-proven clinical images.<sup>16</sup> With the prevalence of smartphones, skin lesion classification with DL could potentially be available directly to patients.<sup>29</sup> DL also has the potential to aid doctors with diagnostic procedures such as colonoscopies by analyzing live videos and highlighting suspicious regions of tissue in real-time to guide the operation.<sup>17</sup> In radiology, Ardila et al. developed a 3D CNN for lung cancer screening with one component identifying regions of interest (ROIs), another component processing the entire image, and a final classification layer combining the outputs of both components.<sup>30</sup> If a prior CT scan is available, the model extracts features from ROIs in both the current and prior CT images. Their model was at least on par with six radiologists and reduced both false positive and false negative rates in some situations. While many such methods were validated on relatively small datasets from a single site, McKinney et al. built a DL model for diagnosing breast cancer from mammograms and evaluated their model on large datasets from the US and the UK.<sup>31</sup> They found that their model had superior performance compared to six radiologists. They also demonstrated that in many cases, they could replace a second reader, which is standard procedure in the UK, with their model's prediction and save 88% of the time of the second reader without sacrificing performance.

Despite these successes, there has been debate about the transparency, interpretability, reproducibility, and robustness of some of these results.<sup>32</sup> Most of these studies are retrospective, single-site, and compare ML performance *post hoc* to human performance rather than evaluating ML models in the way they would be used in the clinic, as a system to assist human decision making. Some recent studies have worked to address these shortcomings to more convincingly demonstrate the benefits of ML in cancer diagnosis. Qian et al. performed a prospective, rather than retrospective, evaluation of a DL model using ultrasound to assess breast cancer.<sup>33</sup> Kim et al. designed a reader study in which radiologists evaluated mammograms either with or without the aid of an ML model trained on mammograms from five institutions in three countries.<sup>34</sup> Radiologists from multiple institutions had superior performance when working in conjunction with ML rather than alone. Hekler et al. had dermatologists and an ML model separately evaluate skin images to detect cancer and then combined those predictions using a decision tree-based ML algorithm called XGBoost to achieve performance superior to either method independently.<sup>35</sup>

Image-based deep learning has also been used in other ways to aid preliminary cancer diagnosis. In Yala et al., a ResNet-18 model was built to triage mammograms by setting a high-sensitivity prediction threshold so that nearly all predicted negative cases were truly negative.<sup>36</sup> In a simulation study, these predicted negative cases were skipped by radiologists, allowing radiologists to only read 80.7% of mammograms while maintaining sensitivity and specificity across all cases. Instead of diagnosing

cases, Xu et al. built a CNN model to segment breast ultrasound images into functional tissues to aid clinicians who interpret and diagnose the images.<sup>37</sup> Cao et al. designed a model that simultaneously diagnoses and grades prostate cancer at the pixel level from multi-parametric MRI, leveraging the power of DL models to move beyond cancer detection alone.<sup>38</sup> A future direction is integrating patient history and pertinent clinical presentation in image-based DL models. Multimodal DL models have become increasingly popular in healthcare applications, given the importance of clinical history in diagnosis. In one instance, Akselrod-Ballin et al. trained a DL model to diagnose breast cancer from mammograms that additionally incorporates information from medical records, finding that it led to improved diagnostic accuracy over models that did not incorporate health records.<sup>39</sup>

### Confirmation by pathology

Pathology samples, typically stained with hematoxylin and eosin (H&E), are assessed by pathologists to confirm a preliminary cancer diagnosis. Due to the large size of digital whole slide images of histopathology, DL models frequently use multiple instance learning (MIL). In MIL, the DL model operates on small image tiles and then aggregates individual tile-level embeddings or predictions into a diagnostic prediction for the whole slide.<sup>40</sup> Campanella et al. used MIL to train a DL model for prostate, breast, and other cancers. The model could allow pathologists to exclude 65–75% of slides while still identifying cancers with 100% sensitivity.<sup>41</sup> This model has the potential to significantly reduce the workload of pathologists, allowing them to spend more time on difficult cases.

As with preliminary diagnosis via non-invasive imaging, rigorous evaluations of DL-based pathology tools using multi-site, prospective trials with DL-assisted pathologists are needed to evaluate the clinical utility of these models. Several recent works have performed studies with at least some of these criteria, showing improved pathologist performance when assisted by DL models that highlight ROIs of the image and/or provide a diagnostic prediction.<sup>42,43</sup>

DL models sometimes predict more than a binary cancer versus no cancer label in order to provide clinicians with additional diagnostic information. For example, in cases of cancer of unknown primary origin, determining an appropriate diagnosis and treatment plan requires inferring the origin of cancer. Lu et al. trained a ResNet-50-based model on H&E images to identify a tumor as primary or metastatic and predict its site of origin across 18 different primary origins, with top-3 prediction accuracy on an external set exceeding 90%.<sup>15</sup> The model incorporated attention after the CNN layers, which identified regions in the slide of high diagnostic relevance and provided a form of human interpretability. Coudray et al. built an Inception v3 CNN model for lung cancer to simultaneously diagnose cancer, determine the tumor subtype of positive cases, and predict the presence of six genetic mutations from H&E-stained images.<sup>18</sup>

### Prognosis and treatment selection

After a cancer diagnosis, physicians and patients are interested in determining the patient's prognosis and selecting the optimal treatment for that patient. Since both prognosis and treatment selection depend on the characteristics of the cancer, many ML methods indirectly aid prognosis and treatment selection

by predicting tumor features such as cancer subtype or mutation status. Other methods directly predict prognosis or guide treatment selection by evaluating or planning potential treatments. Below, we discuss both types of methods.

### **Tumor features**

Prognosis and treatment selection are both informed by a number of tumor features that can be predicted by image-based ML models. For example, ML models have been developed to predict the subtype or grade of a tumor, such as the Gleason grade in prostate cancer,<sup>44</sup> which gives physicians information about patient survival and which treatments might be most effective. Esteva et al. fuse information from both histology slides and clinical data in a DL model that predicts the likelihood of 5- and 10-year metastasis, which can indicate more aggressive disease that requires additional treatment.<sup>45</sup> They pre-trained the image portion of their DL model using a self-supervised technique called momentum contrast, in which the model was trained to identify whether two image tiles were augmented versions of the same tile or were different tiles. Besides tumor subtype, another goal is to predict the genetic characteristics of a tumor, such as microsatellite instability,<sup>46</sup> tumor mutational burden,<sup>47</sup> or whole-genome duplication.<sup>48</sup> Some studies use H&E images to predict gene expression and assess survival-related tumor heterogeneity.<sup>49</sup> Saltz et al. develop a deep learning-based computational stain that identifies tumor-infiltrating lymphocytes whose spatial patterns are correlated with survival.<sup>50</sup> Wang et al. use a 3D CNN to predict EGFR mutation status in lung adenocarcinoma from ROIs selected manually from CT scans, thus providing a non-invasive method of genotyping cancer and informing potential treatments.<sup>51</sup> When biopsy samples are available, it is still more reliable to measure genotypes using molecular methods that we discuss in the next section.

### **Prognosis**

A number of DL models have been developed to predict patient survival from histology slides. Courtiol et al. provide an example of this type of model and workflow for prognosis in mesothelioma.<sup>52</sup> First, they trained a U-Net CNN on several hundred manually annotated histology images to perform tissue segmentation. Next, they divided each patient's whole slide histology image into small image tiles and kept all the tiles that were predicted to contain at least 20% tissue according to the U-Net model. Using transfer learning, they took a ResNet-50 CNN pre-trained on an image recognition task called ImageNet and used it to predict a score for each tile. The 10 highest and 10 lowest scores were passed to a neural network that predicts the patient's survival time. The ResNet-50 model and neural network were trained together on 2,300 slides with a loss function based on the Cox proportional hazards model. They demonstrate that their model significantly outperforms simpler survival prediction models that only use histological type or grade without the image. Bychkov et al. instead predict survival for colorectal cancer using all image tiles by applying an RNN to aggregate the embeddings produced by a CNN model for each tile.<sup>53</sup> In contrast to methods using histology images, Xu et al. take advantage of the fact that radiology is non-invasive and can easily be repeated over time to develop a combined CNN + RNN model that updates its survival predictions over the course of treatment.<sup>54</sup>

### **Response to treatment**

Predicting response to treatment, either prior to or during the early stages of treatment, can aid physicians in selecting the optimal treatment for a patient. Joo et al. developed a multi-modal DL model to predict whether patients would achieve a pathologic complete response after neoadjuvant chemotherapy (NAC) for breast cancer.<sup>55</sup> Their model made predictions by fusing information from two different pretreatment MRIs, each processed with a 3D ResNet model, and clinical information, such as age and HER2 status, processed by a neural network. Gu et al. also aimed to predict response to NAC in breast cancer, but they applied DL models to pairs of ultrasonography images, with one image taken before NAC and the other taken after some, but not all, of the NAC treatments.<sup>56</sup> Through a prospective study, they showed that their model could predict whether a patient would respond to the full course of therapy, indicating that it could be used to alter the course of treatment early for those patients who are predicted not to respond. Tian et al. built a model that extracts features from CT images using a DenseNet CNN and hand-crafted radiomics features, with a neural network classifier processing the concatenation of both sets of features to assess PD-L1 expression in non-small cell lung cancer.<sup>57</sup> This enables a non-invasive prediction of response to anti-PD-1 antibody immunotherapy. Lu et al. found that deep learning could evaluate tumor morphological change in metastatic colorectal cancer from CT scans, which may allow early adjustments during treatment.<sup>58</sup> Notably, this study used an RNN to combine image features extracted by CNNs from CT scans at multiple time points during treatment.

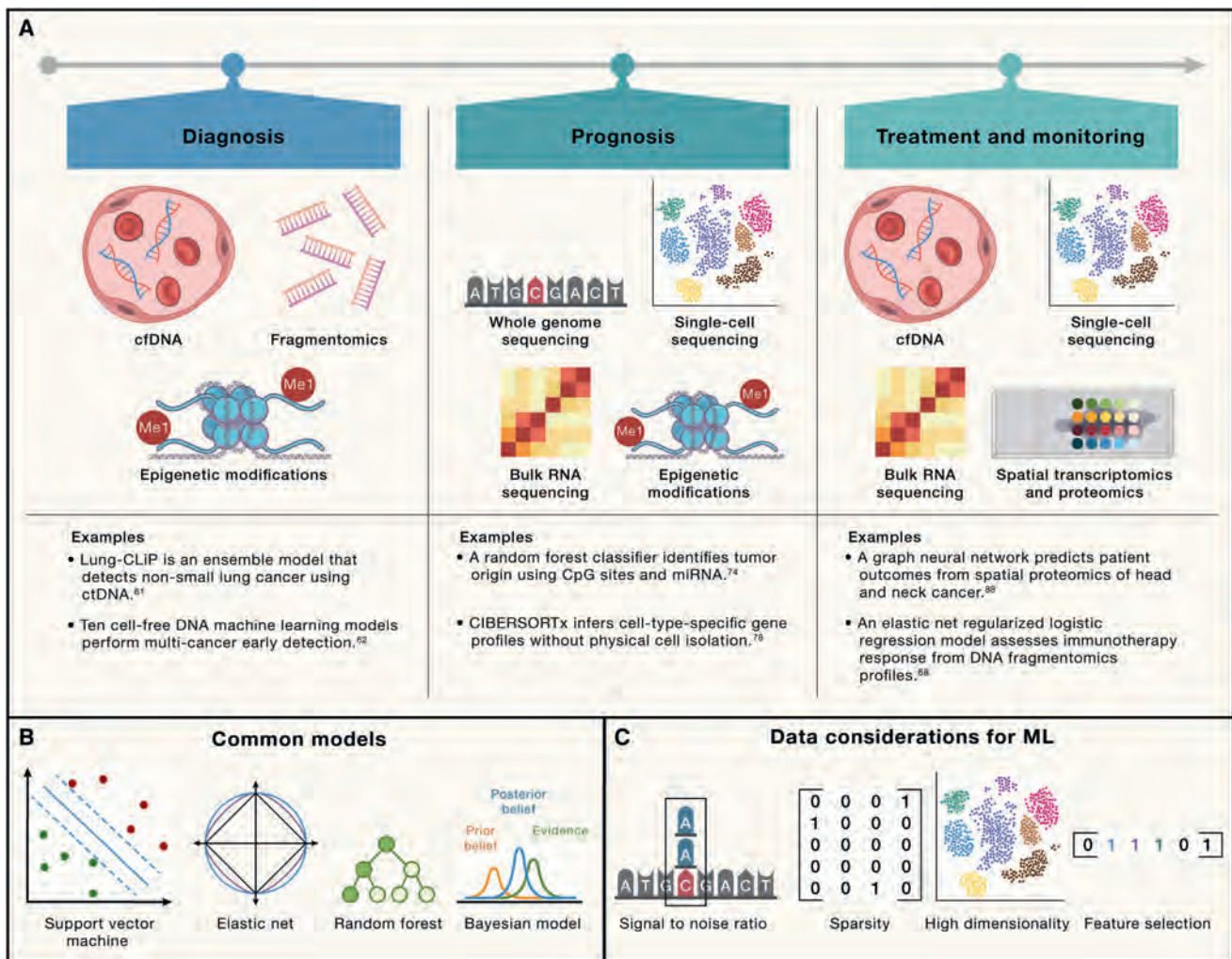
### **Radiotherapy planning**

Planning radiotherapy is a time-consuming process that could benefit from the speed of ML models. McIntosh et al. performed a blinded, head-to-head study of human-generated and ML-generated radiotherapy treatment plans for prostate cancer.<sup>5</sup> ML-generated treatment plans were inferred from the treatment plans of previous patients who were most similar to the current patient according to a learned similarity metric based on features extracted from CT images. In their prospective study of 50 patients, ML-generated plans were selected over human-generated plans 61% of the time while reducing the radiotherapy planning time by 60%, from a median of 118 h to 47 h. Hosny et al. built a U-Net model to segment primary non-small cell lung cancer tumors and involved lymph nodes in CT images, which is a time-consuming step in radiotherapy planning, with validation across eight internal and external clinical sites in multiple countries.<sup>59</sup> In their study, AI assistance led to a 65% reduction in segmentation time and a 32% reduction in variability between clinicians.

## **MACHINE LEARNING FOR MOLECULAR CANCER DIAGNOSIS, PROGNOSIS, AND TREATMENT**

Recent advances in sample processing, genomic sequencing, and molecular technologies have generated rich datasets from solid tumor biopsies and molecular liquid biopsies, which aim to detect circulating cell-free tumor DNA (cfDNA). ML models have played an instrumental role in mapping these datasets to clinical outputs. We first give an overview of liquid biopsy and solid tumor datasets and discuss how their unique





**Figure 3. Machine learning for molecular cancer diagnosis, prognosis, and treatment**

(A) Common molecular datasets for molecular cancer diagnostics include circulating cell-free DNA (cfDNA), methylation status, and fragmentomics. Many molecular datasets for cancer prognosis have been generated from whole-genome sequencing, single-cell transcriptomics, and bulk RNA sequencing of solid tumor biopsies. Utilizing molecular datasets for cancer treatment prediction and selection is a rapidly developing field incorporating foundational molecular technologies and emerging methods such as spatial omics. Example studies are given.

(B) Designs of common ML models for molecular data.

(C) Considerations of molecular data that inform the choice of ML model.

characteristics influence the ML models utilized. We focus our attention on how ML models have been applied for tertiary analysis of genomic datasets. We then give an overview of how ML models have been applied to facilitate liquid biopsy-based and solid tumor-based diagnosis, prognosis, and treatment selection and tumor monitoring. These advancements, summarized in Figure 3, have spurred a rapidly developing field that has garnered tremendous clinical and commercial interest.

### Characteristics and ML models for molecular datasets

Liquid and solid tumor biopsy data sequencing datasets share several characteristics and challenges that guide the design of ML methods. First, dataset size is often limited. Each tumor subtype may only be represented by less than 50 samples.<sup>60</sup> Given the small number of samples per dataset, ML models tend to be

smaller and leverage careful feature engineering and domain expertise.<sup>61</sup> Ongoing initiatives, such as the Circulating Cell-free Genome Atlas (CCGA), which has recruited 15,000 patients from over 140 sites, will provide valuable new resources that are multi-institutional and balanced in patient and clinical demographics.<sup>62</sup>

The small sample challenge of liquid and solid tumor biopsy datasets is amplified by the high-dimensional nature of the data. Thus, applying ML to liquid and solid biopsy datasets requires careful consideration of properly selecting features or aggregating existing features for model training. Additionally, high dimensionality warrants vigilance for overfitting to training data.<sup>63</sup> Here, regularization, which regularizes or pressures model coefficients toward zero in order to encourage less complex and flexible models that are less susceptible to overfitting,

have been helpful in mitigating problems that arise with high-dimensional datasets. Common regularization methods that have been used with molecular datasets include ridge, LASSO, or elastic net.

Molecular datasets can also suffer from a low signal-to-noise ratio stemming from difficulties in determining the veracity of detected variants.<sup>64</sup> Of note, circulating tumor DNA (ctDNA) typically comprises only 5%–10% (in late-stage disease) to less than 0.01%–1.0% (in early-stage disease) of total circulating cell-free DNA.<sup>64</sup> The balance between wide coverage but low sequencing depth versus high sequencing depth of a more limited target is an important factor that affects the signal-to-noise ratio.<sup>65</sup> This tradeoff is further amplified when creating molecular datasets for ML applications. Targeted sequencing panels can reduce noise; however, emerging work has demonstrated that aggregation variants across the genome can improve ML performance. Careful design of training datasets for ML applications can help to mitigate some of the noisy data limitations. Case-control designs—e.g., cases comprising patients with localized non-small cell lung cancer matched with controls of risk-matched adults undergoing annual radiologic screening for lung cancer—are a common strategy to reduce confounders and improve signal.<sup>61</sup>

While DL has become the model of choice for numerous genomic applications, the unique challenges of liquid and solid tumor biopsy data have rendered DL models less directly applicable. Moreover, inductive biases of popular DL architectures (e.g., spatial invariance of CNNs) are less suitable for sequence variants or gene expression. Rather, smaller models such as regularized logistic regression,<sup>61</sup> SVM,<sup>66</sup> random forest classifiers,<sup>67</sup> and elastic nets<sup>68</sup> are commonly used, and they utilize domain expertise to design features.<sup>66</sup>

### Applications of ML models to molecular tumor data

In this section, we review how ML is facilitating the use of molecular data for cancer diagnosis, prognosis, and treatment selection and tumor monitoring (Figure 3).

#### Cancer diagnosis

Early cancer detection is critical for timely interventions that can improve patient outcomes. Liquid biopsy methods utilize detected variants from a targeted sequencing panel to determine the presence of cancer. While detected mutational burden can be predictive, using mutational burden alone can be limited in sensitivity, specificity, and power.<sup>61</sup> Integrating additional variants and genomic features can increase predictive power. ML models have been instrumental in classifying detected variants as pathological, aggregating variants, and identifying variants that are most predictive.

Models such as logistic regression<sup>69</sup> and elastic net<sup>61</sup> have been used to integrate detected variants. For example, Lung-CLiP (Cancer Likelihood in Plasma) employs an ensemble ML classifier using nearest neighbor classifiers, naive Bayes, logistic regression, and decision trees to determine the likelihood that a plasma sample contains lung cancer ctDNA.<sup>61</sup> While detecting variant burden from cfDNA is promising, ascertaining the tissue of origin of ctDNA is more challenging.

DNA methylation sequences have also been pursued as molecular predictors for early cancer detection. Changes to CpG

DNA methylation are one of the earliest molecular aberrations in cancer initiation and offer enhanced capability to infer tissue origin of ctDNA due to the presence of tissue-specific CpG islands. A systematic evaluation of 10 ML classifiers with various data inputs (whole-genome sequencing of cfDNA, targeted cfDNA panels, and DNA methylation) using CCGA found that classifiers that utilized whole-genome methylation sequences had the highest cancer detection sensitivity and best prediction of cancer signal origin.<sup>62</sup>

A central challenge in utilizing methylation sequences is determining which methylation features to select, given that there are 30 million CpG sites that can be methylated or unmethylated. This can be tackled through ML methods that facilitate dimensionality reduction or feature selection. Regularized regression, such as elastic net, has been popular in feature selection for methylation datasets.<sup>70</sup> Maros et al. systematically compared four ML classifiers (random forest, elastic net, SVM, and boosted trees) in combination with post-processing algorithms and found that elastic net delivers the best performance in methylation-based cancer detection and classification.<sup>71</sup> Grail has utilized probability models, such as Bernoulli mixture models, to determine the ranking of positive and negative methylation features likely to distinguish cancer types from one another or non-cancer.<sup>72</sup>

While previous liquid biopsy technologies have primarily utilized cfDNA sequences or methylation status, the fragmentation patterns of cfDNA, also called fragmentomics, can provide additional features to enhance ML cancer detection models. Several studies have found that incorporating fragmentomics into their classifier improved classifier performance.<sup>61,67</sup> Similarly, Jamshidi et al. found that fragment length ML classifiers provided similar sensitivity to a classifier based on genomic alterations.<sup>62</sup> Improved performance could be attributed to additional epigenetic or mechanistic information conveyed by fragmentomic profiles that can increase predictive capability. For example, Esfahani et al. utilized an elastic net model trained on fragmentomics to infer gene expression, classify non-small cell lung cancer, and assess immunotherapy response.<sup>68</sup>

#### Cancer prognosis

While liquid biopsies hold the potential to revolutionize cancer diagnostics, solid tumor molecular analysis is currently more mature and can provide high-resolution molecular and clinical information that can be leveraged to better characterize cancer prognosis.

Advances in exome and whole-genome sequencing and bulk and single-cell transcriptomic technology offer exciting opportunities to characterize tumor origin, stage, and grade, which influence cancer prognosis. Determining tumor origin, particularly for metastatic tumors, is an important aspect of cancer prognosis that molecular ML models can facilitate. Random forest classifiers have been a popular model of choice for predicting tumor origin. For example, Nguyen et al. utilized an ensemble of binary random forest classifiers trained on 6,756 whole-genome-sequenced primary and metastatic tumors that discriminated between 35 cancer types with an overall recall of 90%.<sup>73</sup> Similarly, Tang et al. developed a random forest classifier trained on methylation and miRNA expression data from 17 classes of solid tumors to predict tumor origin.<sup>74</sup> For metastatic tumors,

researchers developed random forest models that perform feature selection and tissue-of-origin classification using gene expression and mutation data.<sup>75</sup> Random forest classifiers are popular due to their ease of interpretability, which provides mechanistic justification of predictions and can facilitate novel biomarker discovery. However, random forest classifiers often require hand-selected features that have relied on patterns of somatic mutations and chromatin state for determining tumor origin. Using a fully connected, feedforward neural network, Jiao et al. determined features correlated with tumor origin and found that passenger mutation regional distribution and mutation type strongly predict tumor origin.<sup>76</sup>

Determining cell-type composition in tumors is critical in assessing cancer prognosis, as it gives insight into the differentiation status, tumor origin, and stage. Several methods have been developed to deconvolve bulk RNA-seq data, a common and cost-effective method to profile solid tumors. Methods such as CIBERSORT use SVMs to deconvolve bulk RNA-seq data to estimate cell-type compositions.<sup>77</sup> CIBERSORTx and CODEFACS have expanded CIBERSORT to deconvolve bulk RNA using nu-support vector regression ( $\nu$ -SVR) analysis and achieve cell-type-specific gene expression without single-cell data.<sup>78,79</sup> While most deconvolution efforts have thus far focused on bulk cellular tissue sources such as tumor specimens, ML deconvolution applications to cell-free nucleic acids are emerging. Indeed, inference of cell types of origin within cell-free RNA (cfRNA) transcriptomes has been achieved using adaptations of CIBERSORTx and  $\nu$ -SVR,<sup>80</sup> as well as using Bayesian cell proportion reconstruction inferred using statistical marginalization.<sup>81</sup>

In addition to DNA mutations and RNA expression, DNA methylation patterns can also differentiate between different cancer types and subtypes. Capper et al. take advantage of this by designing an ML model that can assign central nervous system tumor (CNS) samples to methylation classes that correspond to tumor types based on genome-wide methylation data.<sup>82</sup> Their model consists of a random forest to compute raw scores for the methylation classes followed by a multinomial logistic regression model to calibrate those scores as probabilities of each class. In two prospective analyses, they showed that the methylation predictions perform comparable to or better than histopathological analysis in subtyping some tumors. As an alternative to genomic and transcriptomic methods, Klein et al. used mass spectrometry to analyze epithelial ovarian cancer, and they developed SVM and 1D CNN models that analyze the mass spectrum and predict the histotype of the tumor.<sup>83</sup>

### **Cancer treatment and tumor monitoring**

Selecting cancer treatment, predicting response to treatment, and monitoring tumors after treatment are areas of great promise for ML and genomics. Current treatment selection is determined by clinical guidelines and trials that typically use a handful of clinical features. In contrast, molecular profiles of cancers generate a much larger number of features that can be leveraged to inform cancer treatments. For example, Sammut et al. take a multi-omics approach to predict response to chemotherapy by incorporating clinical, genomic, transcriptomic, pathology, and treatment information into an ensemble model that averages



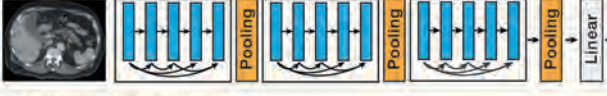
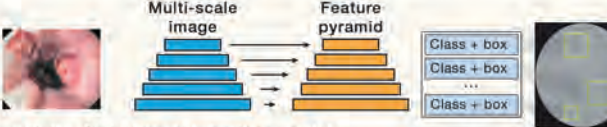
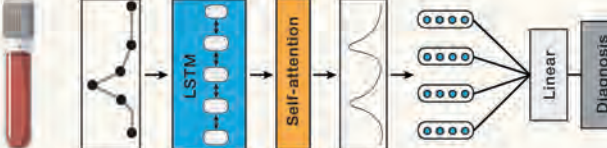
the predictions of logistic regression, SVM, and random forest models.<sup>84</sup> Bayesian models such as the continuous individualized risk index (CIRI), which are adept at handling small datasets and quantifying uncertainty, have been used to model ctDNA dynamics after treatment in diverse cancers.<sup>85</sup> Such approaches can model ctDNA responses associated with outcomes after therapy with immune checkpoint inhibitors for non-small cell lung cancer and predict which patients will achieve durable clinical benefit.<sup>86</sup> New emerging genomic technologies, such as single-cell transcriptomics and spatial transcriptomics, have the potential to revolutionize histopathology characterization of solid tumors. In particular, single-cell transcriptomics can profile the cell composition, which ML models can leverage to predict cancer treatment response and potential resistance.<sup>87</sup> Graph neural networks trained on spatial proteomics can model the tumor microenvironment and predict patient response to cancer treatments.<sup>49,88</sup>

### **REGULATORY APPROVAL OF CANCER ML ALGORITHMS**

The ML algorithms reviewed in the previous sections reflect notable advances in the research landscape. However, before ML algorithms can be deployed on patients, they generally require regulatory approval, which entails more rigorous clinical trials and validation testing than what is presented in published academic work. As such, only a small proportion of ML algorithms end up being deployed on patients. Of those that do, they typically perform well in several predefined tasks like detection and triage settings, and they demonstrate reliability and generalizability across different patient populations.

In the US, most ML algorithms are regulated as medical devices by the Food and Drug Administration (FDA). In the past decade, over 300 AI/ML-enabled medical devices have been approved by the FDA, with over 40% approved since 2020.<sup>89</sup> As an exception to FDA approval, laboratory-developed tests (LDTs) may alternatively receive Clinical Laboratory Improvement Amendments (CLIA) certification by the Centers for Medicare & Medicaid Services (CMS). Certification of such CLIA LDTs generally applies a lower regulatory standard for approval than the FDA.<sup>90</sup> While FDA-approved medical devices are approved for use by medical practitioners, CLIA-certified LDTs are approved for use only by the laboratory for which the certification is granted. LDTs have become increasingly complex and often use ML. The FDA has called for stricter regulation and oversight particularly over higher-risk LDTs,<sup>90</sup> though regulatory changes remain to be implemented. Figure 4 summarizes several examples of regulatory-approved ML medical devices for cancer, including clinical study and ML model details, and Table 1 shows additional examples of approved devices.

The European Union's FDA equivalent, the European Medicines Agency (EMA), operates similarly: cancer-diagnostic AI/ML devices are given a CE mark, which grants approval for sale across the EU and other European countries. However, unlike the US FDA, EMA device approval is decentralized, where individual member countries conduct evaluations, and publicly available information on approvals is sparse. In a comparative analysis of ML devices approved by both the US FDA and

Device description	Clinical study details	Model diagram and description
<p><b>Transpara</b></p> <ul style="list-style-type: none"> <li>Breast cancer mammography detection algorithm</li> <li>FDA approved, 2020</li> </ul>	<ul style="list-style-type: none"> <li>AI-assisted and standalone studies                             <ul style="list-style-type: none"> <li>18 readers</li> <li>240 exams</li> </ul> </li> </ul>	 <p>Key: Feature processing (blue), Feature aggregation (orange), Linear layers (grey), Prediction (dark grey).</p> <ul style="list-style-type: none"> <li>RetinaNet object detection model</li> <li>Outputs image and lesion scores</li> </ul>
<p><b>Paige Prostate</b></p> <ul style="list-style-type: none"> <li>Prostate pathology cancer diagnostic algorithm</li> <li>FDA approved, 2019</li> </ul>	<ul style="list-style-type: none"> <li>Standalone analytical testing                             <ul style="list-style-type: none"> <li>847 slides</li> </ul> </li> <li>AI-assisted study                             <ul style="list-style-type: none"> <li>527 slides</li> <li>16 pathologists</li> </ul> </li> </ul>	 <ul style="list-style-type: none"> <li>ResNet-34 CNN feature extractor</li> <li>RNN for score prediction</li> <li>Multiple instance learning</li> </ul>
<p><b>Optellum</b></p> <ul style="list-style-type: none"> <li>Lung CT cancer nodule detection algorithm</li> <li>FDA approved, 2021</li> </ul>	<ul style="list-style-type: none"> <li>AI-assisted and standalone studies                             <ul style="list-style-type: none"> <li>300 subjects</li> <li>12 readers</li> </ul> </li> </ul>	 <ul style="list-style-type: none"> <li>DenseNet CNN classifier</li> </ul>
<p><b>GI Genius</b></p> <ul style="list-style-type: none"> <li>Lesion detection for endoscopy video</li> <li>FDA approved, 2021</li> </ul>	<ul style="list-style-type: none"> <li>Standalone study                             <ul style="list-style-type: none"> <li>150 videos</li> <li>338 lesions</li> </ul> </li> </ul>	 <ul style="list-style-type: none"> <li>RetinaNet object detection model</li> <li>Video frames are individually processed</li> </ul>
<p><b>InterVenn GLORI</b></p> <ul style="list-style-type: none"> <li>Lab developed test for ovarian cancer diagnosis</li> <li>CLIA certified, 2021</li> </ul>	<ul style="list-style-type: none"> <li>Prospective observational study                             <ul style="list-style-type: none"> <li>1,200 participants</li> </ul> </li> </ul>	 <ul style="list-style-type: none"> <li>LSTM model for signal processing</li> <li>Regression model for score prediction</li> </ul>

**Figure 4. Regulatory approval of cancer ML algorithms**

Examples of ML medical devices for cancer that have received regulatory approval, including Transpara,<sup>91</sup> Paige Prostate,<sup>38</sup> Optellum,<sup>92</sup> GI Genius,<sup>93</sup> and InterVenn GLORI.<sup>94,95</sup> Clinical study details are based on information available in published works and registered clinical trials. Model details are based on publications by device developers. Image sources: mammography,<sup>19</sup> CT,<sup>20</sup> histology,<sup>11</sup> endoscopy.<sup>96</sup>

EMA, most devices first received approval in Europe, suggesting a potentially lower regulatory bar compared to the US.<sup>105</sup>

**Imaging-based algorithms**

Imaging-based algorithms comprise over 70% of all FDA-approved AI/ML devices.<sup>106</sup> Of these, radiology applications are the most abundant. Pre-diagnosis algorithms like WRDensity and Densitas use CNN architectures like ResNet<sup>102</sup> to provide breast density category predictions for mammograms. AI-Rad Companion and Quantib Prostate use CNN-based networks<sup>100</sup>

like U-Net to perform automated segmentation, density calculation, and volume estimation of the prostate gland. Computer-aided triage devices like Saige-Q<sup>107</sup> and CmTriage use CNN classification algorithms to mark a subset of mammogram cases as suspicious to aid radiologists in workload prioritization. Computer-aided detection/diagnosis devices provide more information by identifying and scoring regions of interest in each image. Examples of breast cancer devices include Lunit Insight, which draws heatmaps (using convolutional layers) with probability percentages over suspicious regions in a mammogram,<sup>34</sup> and

**Table 1. Additional examples of regulatory-approved cancer diagnostic devices**

Approval type (#)	Date approved	Device name	Description	Type of AI/ML	Clinical study details
FDA (P170019)	2017	FoundationOne CDx	Microsatellite instability and tumor mutational burden solid tumor tests	Probit model for level of detection <sup>98</sup>	Prospective observational studies (1,400 participants)
CLIA certification	2017	Signatera	LDT for ctDNA-based cancer recurrence test	Cox proportional hazards model <sup>97</sup>	Prospective observational studies (2,000 participants), still recruiting
FDA (K173839)	2017	The Cancer Genetics Tissue of Origin Test	Tissue of origin genetic test	Normalization, classification, and correlation algorithms <sup>99</sup>	Analytical testing only
FDA (K183271)	2019	AI RAD Companion (Pulmonary)	Lung nodule segmentation	FCOS CNN object detection network <sup>100</sup>	>4,500 cases, standalone study only, reader-annotated ground truth
FDA (K183285)	2019	CmTriage	Breast cancer triage	CNN <sup>101</sup>	1,255 exams, standalone study only, biopsy-proven ground truth
FDA (K200595)	2020	CellaVision DC-1	Blood cell counter	CNN	Analytical and clinical testing (598 samples) comparing to predicate device
FDA (K201232)	2020	Limbus Contour	Radiation treatment planning	U-Net CNN <sup>103</sup>	Benchtop testing only
FDA (K193229)	2020	Transpara	Breast cancer detection	VGG-16 CNN and gradient boosting trees <sup>91</sup>	240 exams, AI-assisted (18 readers) and standalone studies, ground truth unclear
FDA (K202013)	2020	WRDensity	Breast cancer density	Resnet-34 CNN <sup>102</sup>	871 exams, standalone study only, consensus ground truth
FDA (K211951)	2021	GI Genius	GI lesion detection	CNN object detection network <sup>93</sup>	Standalone study only (150 videos with 338 lesions)
CLIA certification	2021	Grail Galleri	Multi-cancer early detection test	Various ML models (logistic/lasso regression, Markov chains, random forest) <sup>104</sup>	Prospective observational and interventional studies (>130K participants)
CLIA certification	2021	InterVenn GLORI	LDT for ovarian cancer diagnosis	Regression models and RNNs <sup>94,95</sup>	Prospective observational study (1,200 participants), ground truth by imaging
FDA (DEN200080)	2021	Paige Prostate	Prostate pathology cancer detection	ResNet-34 CNN + RNN (multiple-instance/weakly supervised learning) <sup>41</sup>	Standalone analytical testing on 847 whole slide images (WSIs) and AI-assisted study on 527 WSIs with 16 pathologists, consensus ground truth
FDA (K202300)	2021	Optellum Virtual Nodule Clinic	Lung nodule diagnosis	DenseNet CNN <sup>92</sup>	300 subjects, AI-assisted (12 readers) and standalone studies, ground truth unclear

MammoScreen, which uses a RetinaNet CNN architecture to draw a bounding polygon over potential lesions along with the predicted lesion type and risk score out of ten.<sup>108</sup> Another example is Optellum Virtual Nodule Clinic, a lung cancer algorithm for CT images that uses a DenseNet architecture to output malignancy prediction scores for user-selected regions of interest.<sup>92</sup>

Imaging ML has more recently expanded outside of radiology as well. Paige Prostate is an FDA-approved prostate pathology algorithm, based on the work of Campanella et al.,<sup>41</sup> that uses CNNs and RNNs to diagnose prostate cancer from biopsy slides.<sup>109</sup> Other prostate CLIA-certified pathology ML tests include DeepDx Prostate, which uses semantic segmentation CNNs, and Galen Prostate, which uses multiscale CNNs and

gradient boosting classifiers for automated Gleason scoring.<sup>110</sup> GI Genius, an FDA-approved device for polyp detection in endoscopy videos, uses a CNN on individual video frames to produce bounding boxes over suspicious lesions.<sup>93</sup>

Skin cancer is a promising yet challenging domain. Nevisense, currently the only skin cancer AI device on the market, is a device that works by measuring electrical impedance across a potentially abnormal skin lesion. On the horizon, 3Derm has received FDA breakthrough designation for autonomous detection of skin cancers, which is a fast-track process that signals possible future approval. In the EU, several skin AI devices have already received CE mark approval (TeleSkin and SkinVision), but their efficacy has been questioned by independent validation studies.<sup>111</sup>

Several devices have been approved for post-diagnosis decision making; for instance, Limbus Counter and Ethos are both devices that use segmentation CNNs like U-Net to draw contours of organ structures for radiation treatment planning.<sup>112</sup>

### Molecular-based algorithms

Most molecular-based algorithms are focused on diagnostic applications in blood samples. FDA-approved cell counting devices like CellaVision and Sight OLO use CNNs to characterize and count white blood cells, red blood cells, and platelets in blood samples and are intended for use by lab technicians.<sup>113</sup> CellSearch uses computer vision algorithms to characterize the morphology of circulating tumor cells in metastatic breast, colorectal, or prostate cancer patients. The Cancer Genetics Tissue of Origin Test is an RNA-based diagnostic algorithm for aiding clinicians in determining the tissue of origin for tumors. Exact Science's Cologuard is a colorectal cancer genomics test that relies on mathematical algorithms to produce risk scores.

Liquid biopsy tests are the most common type of ML-enabled diagnostics performed by CLIA-certified laboratories. LungLife AI's LungLB is a liquid biopsy test that uses a signal-binning algorithm to confirm suspicious lung nodules in CT scans. Galleri is a liquid biopsy test that uses various ML regression and classification models<sup>72</sup> for early detection of multiple cancers and has received FDA breakthrough designation but not approval. InterVenn has CLIA certification for two products: GLORI is a glycoproteomic liquid biopsy test that utilizes neural networks and logistic regression models for ovarian cancer diagnosis, and DAWN IO is a test that uses tree-based methods and ensemble classifiers for assessing melanoma therapy.<sup>94</sup> Other genomics tests that are not on the market but are in ongoing large clinical trials include Freenome's Multinomics, a cell-free biomarker patterns blood test using SVM,<sup>60</sup> and Exact Science's multi-cancer early detection blood test.

### Clinical studies evaluating cancer ML algorithms

The types of clinical studies vary depending on the regulatory pathway a device is approved by. For FDA approval, devices must demonstrate evidence of clinical safety and effectiveness for use on patients. Clinical evidence is typically produced via AI-assisted studies and/or standalone studies. AI-assisted studies compare clinicians using AI in diagnostic decision making with those not using AI. In these studies, ground truthing is typically determined by the consensus of several specialists' interpretations. Readers are selected across varying degrees of specialty (generalist versus board certified). Standalone studies provide another form of clinical evidence: the performance of the AI alone is assessed with reference to a reader consensus ground truth, and the metric is compared to the average clinical reader's performance or a standard. In both types of studies, evaluation studies are typically enriched with cancer cases relative to the population incidence rate.

As an example, Transpara, a breast cancer detection algorithm that received FDA approval in 2018, reported clinical evidence from an AI-assisted study and a standalone comparison. Transpara draws regions of interest around suspicious lesions in a mammogram and outputs a score indicating the likelihood of cancer in the image. In the reader study, fourteen board-certified

radiologists read mammograms once with the aid of AI and once without, with a one-month washout period in between. The evaluation dataset consisted of 240 total mammogram studies, with 100 cancer exams, 40 false positive recalls from screening, and 100 normal exams. The primary endpoint was the superiority of performance with AI versus without. Secondary analyses included a superior performance with AI on lesion subtypes and average reading time saved by radiologists. The standalone study compared the AI's performance with the average performance of the fourteen radiologists. In the AI-assisted study, the radiologists' performance improved from 0.866 AUC without AI assistance to 0.886 with AI assistance. In the standalone study, the AI achieved an AUC of 0.887 versus the average clinical reader's performance of 0.866 AUC.

For molecular-based ML device approvals, analytical testing is often conducted in addition to clinical testing. For instance, CellaVision DC-1's FDA evaluation provided evidence demonstrating analytical precision via repeatability (measurements under the same conditions are consistent) and reproducibility (measurements under different conditions are consistent). The clinical testing compared measurements on patient samples with the approved predicate device. Other analytical validation characteristics include accuracy and specificity.

CLIA certifications are less transparent in their evaluation standards compared to the FDA (i.e., no publicly available summaries) but are generally limited to ensuring the analytical validity of lab capabilities. In addition to CLIA certification, most commercially available LDTs have undergone clinical trial validations that are registered with ClinicalTrials.gov. These studies tend to be prospective and larger in scope than FDA-approved device counterparts, which have a median participant size of 300.<sup>89</sup> For instance, Grail's Galleri has ongoing clinical trials with over 130,000 participants across multiple settings and countries. Intervenn's GLORI test enrolled 1,200 patients in its clinical trial. Primary endpoints are similar to FDA evaluations and include AUC, sensitivity, specificity, positive predictive value, and negative predictive value.

## DISCUSSION

ML is increasingly important in cancer detection, prognosis, and treatment planning. However, the reliability and trust of ML algorithms have lagged behind the pace of technical development. In this section, we discuss some key challenges that ML faces on the path to the clinic, including disparate regulatory standards, stringent criteria for meaningful model evaluation, and barriers to adoption by doctors and hospitals. We then discuss how ML methods differ when applied to various cancer data modalities, and we conclude by highlighting some exciting recent developments in both biomedical and ML technologies that illustrate the potential of ML to transform clinical oncology.

### Regulatory standards

Disparate regulatory standards in the US and internationally can lead to under-regulation and mistrust of ML algorithms.<sup>114</sup> Within the US, the FDA has historically deferred the regulation of LDTs to CMS. Whereas CMS typically focuses only on analytical validity (i.e., precision, sensitivity, and accuracy of measuring

molecular quantities), the FDA places additional emphasis on clinical validity (whether the test accurately identifies the relevant disease in patients). As LDTs today increasingly provide diagnostic predictions and involve ML-based algorithms, demonstrating that cancer diagnostic tests truly achieve the desired clinical outcomes is necessary for ensuring their trustworthiness and reliability to doctors and patients.

Discrepancies in regulation internationally contribute an additional risk to the trustworthiness of medical ML algorithms. A study of medical devices approved in both the US and EU revealed that devices that gained CE mark approval first in the EU were three times more likely to be recalled due to safety concerns than devices that received US FDA approval first.<sup>115</sup> A key difference is that in the US, the FDA requires clinical evaluation prior to approval; in the EU, clinical evaluation is only required after approval as a post-market follow-up study.<sup>116</sup> In effect, the CE mark system incentivizes faster adoption of ML into the clinic but at the risk of prematurely approving devices that may pose potential harm to patients.

### Limitations of ML model evaluations

The lack of high-quality, diverse evaluations hinders the ability to assess true algorithm performance in patient populations. One factor is the lack of gold-standard test datasets—on-site validations are difficult and patient data are hard to obtain, in part due to privacy concerns and restrictive data use agreements.<sup>117</sup> A well-documented phenomenon of ML models is that they can learn spurious correlations present in device types and demographics,<sup>89</sup> resulting in biased performance when evaluated on different patient populations. Additionally, evaluation test sets are often enriched with positive cases, which can yield imbalanced comparisons.

### Metrics

Medical AI studies often use proxy metrics for clinical endpoints, which may generate misleading conclusions. For instance, AUC summarizes model performance across all possible operating points, which is not informative of how an algorithm will perform when deployed at a particular threshold. Algorithms that show an AUC improvement or exceed a certain AUC value (e.g., >0.95 in some FDA-approved devices) may perform differently in real-world populations.<sup>118</sup> Fixed-threshold metrics like sensitivity and specificity should reflect the relevant clinical task at hand; for instance, a diagnostic algorithm may be optimized for minimizing missed cancers but should also consider the additional burden to patients caused by false positives (i.e., invasive testing and stress).

### Clinical trials and monitoring

Prospective trials are also important to measure appropriate clinical outcomes, rather than a simple comparison to stand-alone references. For example, if an ML device is to be used as a clinical diagnostic aid, then it should be evaluated by comparing clinician performance with and without the device rather than evaluating the device's predictions in isolation.<sup>119</sup> Randomizing patient cohorts can minimize biases in selecting test populations. Also, prospective trials can capture human-AI interactions that occur after deployment.<sup>120</sup> Continuous performance monitoring of ML algorithms after approval and post-market surveillance mechanisms are necessary to ensure that

the purported clinical benefits of ML hold up under various distribution shifts.<sup>121</sup> As a case study, earlier-generation computer-aided detection software for mammography was approved by the FDA in 1998 and widely adopted in part because of Medicare and Medicaid reimbursements. However, a large observational study by Lehman et al. on mammograms from 2003 to 2009 found that CAD software had failed to improve the diagnostic accuracy of mammography.<sup>122</sup> This was due in part to changes in radiologists' behavior, with increased familiarity with the ML over time.<sup>123</sup> Moreover, the original evaluation data included older traditional film mammograms, which have since been phased out. As such, reproducibility and transparency are essential for building trust in the outcomes of validation studies.<sup>32</sup>

### Interpreting ML models

Interpretability is a common challenge for ML. One important reason is that most models do not explicitly identify causal features but instead rely on correlating input features with outcomes. As such, models may accurately identify phenotypes but rely on spurious confounders present in the data and present misleading conclusions.<sup>124</sup> Nonetheless, interpretability methods can still be useful for explaining how an ML model makes its predictions, which is important for building trust with clinicians and providing additional diagnostic insight beyond the prediction alone.<sup>125</sup> Interpretability methods can either be applied *post hoc* to extract explanations from trained models, or they can be incorporated into the model design so that the model learns to simultaneously produce explanations and predictions. Examples of *post hoc* interpretability techniques include using the ML model to generate heatmaps over the input<sup>33</sup> and clustering the inputs into interpretable groups based on the ML model's embedding of the input.<sup>126</sup> As an example of a model with explainability built into its design, Zhang et al. created an ML model that learns to generate explanations in natural language for its predictions during training.<sup>127</sup> *Post hoc* methods are convenient because they can be applied to most models without requiring specialized training, but models with interpretability built in may provide more reliable explanations for what the model is doing.<sup>125</sup> Models that output a probability or range of scores (e.g., from 1 to 10) should be carefully designed and calibrated to user expectations.<sup>128</sup>

### Challenges to adoption

While most academic research has been focused on improvements in the diagnostic accuracy of ML algorithms, many of the driving factors for real-world clinical ML adoption fall outside of solely technical progress. Interoperability and integration with existing electronic health records and image storage systems is a significant barrier to adoption by hospital systems.<sup>129</sup> Clinicians may not trust or understand ML algorithm decisions and outputs. Developers must effectively communicate the economic value of their ML algorithms to hospital decision makers and overcome organizational inertia. Finally, patients and clinicians should also understand the benefits and risks of using ML in decision making.<sup>130</sup>

### Different data modalities require different ML techniques

Imaging and molecular data are the two most common data modalities in cancer diagnostics. However, in practice, they require

very different ML approaches due to fundamental differences in the problems each data type presents. Imaging-based tasks typically involve a needle-in-the-haystack problem, where small features associated with cancer are present in a large image space. CNNs are highly effective and have become ubiquitous because they are able to efficiently learn from large amounts of available data, and they can extract spatially distinct hierarchies of features present in an image.

Molecular data, on the other hand, tend to be highly structured and have features that correspond to distinct biological measurements (i.e., DNA sequences). A primary hurdle in analysis is the high dimensionality of biological features and the inherent sparsity present in the data. Here, ML regularization techniques like LASSO regression are used, as well as dimensionality reduction techniques like PCA for selecting salient biomarkers. Finally, statistical ML models like logistic regression and decision trees are used to pick optimal thresholds and minimal levels of detection that correspond to a clinically meaningful presence of disease.

### Future developments

New biomedical and ML technologies are rapidly emerging that will change the way ML is applied to cancer diagnostics and may significantly improve the predictive power and clinical usefulness of these models.

### Biomedical data

Biomedical advances are enabling physicians to obtain increasingly detailed medical data about patients. In pathology, new multiplexed proteomics technologies like CODEX<sup>131</sup> allow staining for 40–100 proteins simultaneously, providing a much more detailed view of the cellular and proteomic composition of tissues than traditional staining techniques like H&E staining and immunofluorescence. Similarly, spatial transcriptomics<sup>132</sup> provides a view of the spatial distribution of RNA transcripts across a pathology sample, thereby incorporating another form of omics data into images. Sequencing data from the tumor microbiome might serve as a diagnostic tool for oncology as scientists learn more about the role of bacteria in cancer.<sup>133</sup> Data from the immune system, such as T cell receptor sequences, can also provide diagnostic clues for cancer based on the body's response to tumors.<sup>134</sup> ML methods that use these new sources of data may be able to make more accurate and specific predictions.

### Integrating imaging and omics

Imaging and molecular data often provide complementary information about a patient's cancer, so integrating these two data sources can improve ML predictions for diagnostics, prognosis, and treatment. One method of combining the two is through biomedical technologies such as CODEX and spatial transcriptomics, which overlay spatially resolved proteomics and transcriptomics data on images, allowing models to process omics data in image form.<sup>49,88,135</sup> Another promising direction is the development of multimodal models, which fuse multiple ML models to combine information across several data types (images, genomics, clinical records, etc.) to make better predictions.<sup>2</sup> Multimodal models can have a more holistic view of each patient and can combine multiple weak signals into a strong signal that can better inform the patient's diagnosis or optimal

treatment. For example, Vanguri et al. predict response to PD-(L)1 blockade in patients with non-small cell lung cancer using a multimodal model that combines medical imaging, histopathologic, and genomic features and outperforms unimodal models.<sup>136</sup> Although there are many challenges to developing multimodal models, such as linking data across modalities and handling patients with incomplete data, these models may prove to be very powerful because they can reason across multiple sources of information, just as physicians do.

### ML methodology

New ML models have emerged that improve upon the standard deep learning architectures, such as CNNs, that are commonly used in cancer diagnostics. Several such models have demonstrated clear improvements in predictive accuracy. One of the best examples is the transformer,<sup>137</sup> which was originally designed for natural language processing. Transformers have since been modified and applied to pathology images.<sup>138</sup> Another trend is to re-envision image-based data as a graph and apply GNNs. For example, Wu et al. convert images of tissue samples into graphs of cells, where each cell is a node in the graph and neighboring cells have edges connecting them.<sup>88</sup> GNNs applied to these graphs can make diagnostic and prognostic predictions that may be more robust against visual artifacts and more sensitive to the interconnections between cells than image-based predictions. Instead of using new ML models, another option is to improve the performance of existing ML models by performing data augmentation with generative ML models, which learn to synthesize new data that look similar to the real training data.<sup>139</sup> Generative models are also useful for translating between data formats such as generating text reports from medical images.<sup>140</sup>

The technological advancements discussed in this Review illustrate the exciting potential of ML to leverage the latest biomedical data to transform the field of clinical oncology. As ML methods are further improved and carefully validated with appropriate monitoring and regulatory oversight, they may soon see wide-scale clinical adoption to improve cancer care for patients.

### ACKNOWLEDGMENTS

K.S. is supported by a Knight-Hennessy Scholarship. A.Z. is supported by the National Institutes of Health grant F30HL156478. E.W. is supported by a Stanford Bio-X SIGF Fellowship. J.Z. is supported by a Chan-Zuckerberg Investigator Award.

### DECLARATION OF INTERESTS

A.A.A. is an advisor to Celgene, Chugai, Genentech, Gilead, Janssen, Pharmaclics, and Roche. E.W. is a shareholder of RadNet, Inc. J.Z. is an advisor to Adela, Enable Medicine, and InterVenn Biosciences.

### REFERENCES

1. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17.
2. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J., and Shah, S.P. (2022). Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* 22, 114–126.



3. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., and Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510.
4. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., and Madabhushi, A. (2019). Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* 16, 703–715.
5. McIntosh, C., Conroy, L., Tjong, M.C., Craig, T., Bayley, A., Catton, C., Gospodarowicz, M., Helou, J., Isfahanian, N., Kong, V., et al. (2021). Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.* 27, 999–1005.
6. Bera, K., Braman, N., Gupta, A., Velcheti, V., and Madabhushi, A. (2022). Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat. Rev. Clin. Oncol.* 19, 132–146.
7. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., and Kather, J.N. (2022). Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* 3, 1026–1038.
8. Heitzer, E., Haque, I.S., Roberts, C.E.S., and Speicher, M.R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat. Rev. Genet.* 20, 71–88.
9. Esposito, M., Ganesan, S., and Kang, Y. (2021). Emerging strategies for treating metastasis. *Nat. Cancer* 2, 258–270.
10. Kwong, G.A., Ghosh, S., Gamboa, L., Patriotis, C., Srivastava, S., and Bhatia, S.N. (2021). Synthetic biomarkers: a twenty-first century path to early cancer detection. *Nat. Rev. Cancer* 21, 655–668.
11. Häggström M. Histology of postmenopausal myometrium, low magnification [Internet]. Wikimedia Commons. Available from: [https://commons.wikimedia.org/wiki/File:Histology\\_of\\_postmenopausal\\_myometrium\\_low\\_magnification.jpg](https://commons.wikimedia.org/wiki/File:Histology_of_postmenopausal_myometrium_low_magnification.jpg)
12. Levine, A.B., Schlosser, C., Grewal, J., Coope, R., Jones, S.J.M., and Yip, S. (2019). Rise of the machines: advances in deep learning for cancer diagnosis. *Trends Cancer* 5, 157–169.
13. Lu, M.T., Raghu, V.K., Mayrhofer, T., Aerts, H.J., and Hoffmann, U. (2020). Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Ann. Intern. Med.* 173, 704–713.
14. Varghese, B., Chen, F., Hwang, D., Palmer, S.L., De Castro Abreu, A.L., Ukimura, O., Aron, M., Aron, M., Gill, I., Duddalwar, V., and Pandey, G. (2020). Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA (Association for Computing Machinery), pp. 1–10. (BCB '20).
15. Lu, M.Y., Chen, T.Y., Williamson, D.F.K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. *Nature* 594, 106–110.
16. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118.
17. Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., Takamaru, H., Sakamoto, T., Sese, J., Kuchiba, A., et al. (2019). Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci. Rep.* 9, 14465.
18. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018 Oct). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* 24, 1559–1567.
19. Mammogram - Normal [Internet]. National Cancer Institute Visuals Online. Available from: <https://visualsonline.cancer.gov/details.cfm?imageid=9405>
20. Häggström M. CT of cholecystitis [Internet]. Wikimedia Commons. Available from: [https://commons.wikimedia.org/wiki/File:CT\\_of\\_cholecystitis.jpg](https://commons.wikimedia.org/wiki/File:CT_of_cholecystitis.jpg)
21. Lehman, C.D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., and Barzilay, R. (2019). Mammographic breast density assessment using deep learning: clinical implementation. *Radiology* 290, 52–58.
22. Dontchos, B.N., Yala, A., Barzilay, R., Xiang, J., and Lehman, C.D. (2021 Apr). External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. *Acad. Radiol.* 28, 475–480.
23. Arefan, D., Mohamed, A.A., Berg, W.A., Zuley, M.L., Sumkin, J.H., and Wu, S. (2020 Jan). Deep learning modeling using normal mammograms for predicting breast cancer risk. *Med. Phys.* 47, 110–118.
24. Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., and Strand, F. (2020 Feb). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 294, 265–272.
25. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Smith, K., Wan, Y.L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* 13, eaba4373.
26. Yala, A., Mikhael, P.G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Lehman, C.D., et al. (2022). Multi-institutional validation of a mammography-based breast cancer risk model. *J. Clin. Oncol.* 40, 1732–1740.
27. Ha, R., Chang, P., Karcich, J., Mutasa, S., Pascual Van Sant, E., Liu, M.Z., and Jambawalikar, S. (2019). Convolutional neural network based breast cancer risk stratification using a mammographic dataset. *Acad. Radiol.* 26, 544–549.
28. Yala, A., Mikhael, P.G., Lehman, C., Lin, G., Strand, F., Wan, Y.L., Hughes, K., Satuluru, S., Kim, T., Banerjee, I., et al. (2022). Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.* 28, 136–143.
29. Dai, X., Spasić, I., Meyer, B., Chapman, S., and Andres, F. (2019). Machine learning on mobile: an on-device inference app for skin cancer detection. In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, pp. 301–305.
30. Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* 25, 954–961.
31. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94.
32. Haibe-Kains, B., Adam, G.A., Hosny, A., Khodakarami, F., Massive Analysis Quality Control MAQC Society Board of Directors, Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature* 586, E14–E16.
33. Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., et al. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat. Biomed. Eng.* 5, 522–532.
34. Kim, H.E., Kim, H.H., Han, B.K., Kim, K.H., Han, K., Nam, H., Lee, E.H., and Kim, E.K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multi-reader study. *Lancet. Digit. Health* 2, e138–e148.
35. Hekler, A., Utikal, J.S., Enk, A.H., Hauschild, A., Weichenthal, M., Maron, R.C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *Eur. J. Cancer* 120, 114–121.
36. Yala, A., Schuster, T., Miles, R., Barzilay, R., and Lehman, C. (2019). A deep learning model to triage screening mammograms: a simulation study. *Radiology* 293, 38–46.

37. Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., and Carson, P.L. (2019). Medical breast ultrasound image segmentation by machine learning. *Ultrasonics* *97*, 1–9.
38. Cao, R., Mohammadian Bajgiran, A., Afshari Mirak, S., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., and Sung, K. (2019). Joint prostate cancer detection and gleason score prediction in mp-MRI via FocalNet. *IEEE Trans. Med. Imaging* *38*, 2496–2506.
39. Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., et al. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* *292*, 331–342.
40. Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A.H. (2016). Deep Learning for Identifying Metastatic Breast Cancer. Preprint at arXiv. <http://arxiv.org/abs/1606.05718>.
41. Campanella, G., Hanna, M.G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* *25*, 1301–1309.
42. Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al. (2020). Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat. Commun.* *11*, 4294.
43. Steiner, D.F., MacDonald, R., Liu, Y., Truszkowski, P., Hipp, J.D., Gammage, C., Thng, F., Peng, L., and Stumpe, M.C. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am. J. Surg. Pathol.* *42*, 1636–1646.
44. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol.* *21*, 233–241.
45. Esteve, A., Feng, J., van der Wal, D., Huang, S.C., Simko, J.P., DeVries, S., Chen, E., Schaeffer, E.M., Morgan, T.M., Sun, Y., et al. (2022). Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digit. Med.* *5*, 71.
46. Kather, J.N., Pearson, A.T., Halama, N., Jäger, D., Krause, J., Loosen, S.H., Marx, A., Boor, P., Tacke, F., Neumann, U.P., et al. (2019). Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* *25*, 1054–1056.
47. Jain, M.S., and Massoud, T.F. (2020). Predicting tumour mutational burden from histopathological images using multiscale deep learning. *Nat. Mach. Intell.* *2*, 356–362.
48. Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., and Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* *1*, 800–810.
49. He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., Maaskola, J., Lundeberg, J., and Zou, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* *4*, 827–834.
50. Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K.R., Zhao, T., Batiste, R., et al. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* *23*, 181–193.e7.
51. Wang, S., Shi, J., Ye, Z., Dong, D., Yu, D., Zhou, M., Liu, Y., Gevaert, O., Wang, K., Zhu, Y., et al. (2019). Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur. Respir. J.* *53*, 1800986.
52. Courtiol, P., Maussion, C., Moarii, M., Pronier, E., Pilcer, S., Sefta, M., Manceron, P., Toldo, S., Zaslavskiy, M., Le Stang, N., et al. (2019). Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* *25*, 1519–1525.
53. Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., and Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* *8*, 3395.
54. Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R.H., and Aerts, H.J.W.L. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin. Cancer Res.* *25*, 3266–3275.
55. Joo, S., Ko, E.S., Kwon, S., Jeon, E., Jung, H., Kim, J.Y., Chung, M.J., and Im, Y.H. (2021). Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci. Rep.* *11*, 18800.
56. Gu, J., Tong, T., He, C., Xu, M., Yang, X., Tian, J., Jiang, T., and Wang, K. (2022). Deep learning radiomics of ultrasonography can predict response to neoadjuvant chemotherapy in breast cancer at an early stage of treatment: a prospective study. *Eur. Radiol.* *32*, 2099–2109.
57. Tian, P., He, B., Mu, W., Liu, K., Liu, L., Zeng, H., Liu, Y., Jiang, L., Zhou, P., Huang, Z., et al. (2021). Assessing PD-L1 expression in non-small cell lung cancer and predicting responses to immune checkpoint inhibitors using deep learning on computed tomography images. *Theranostics* *11*, 2098–2107.
58. Lu, L., Dercle, L., Zhao, B., and Schwartz, L.H. (2021). Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. *Nat. Commun.* *12*, 6654.
59. Hosny, A., Bitterman, D.S., Guthrie, C.V., Qian, J.M., Roberts, H., Perni, S., Saraf, A., Peng, L.C., Pashtan, I., Ye, Z., et al. (2022). Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet. Digit. Health* *4*, e657–e666.
60. Wan, N., Weinberg, D., Liu, T.Y., Niehaus, K., Ariazi, E.A., Delubac, D., Kannan, A., White, B., Bailey, M., Bertin, M., et al. (2019). Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* *19*, 832.
61. Chabon, J.J., Hamilton, E.G., Kurtz, D.M., Esfahani, M.S., Moding, E.J., Stehr, H., Schroers-Martin, J., Nabet, B.Y., Chen, B., Chaudhuri, A.A., et al. (2020). Integrating genomic features for non-invasive early lung cancer detection. *Nature* *580*, 245–251.
62. Jamshidi, A., Liu, M.C., Klein, E.A., Venn, O., Hubbell, E., Beausang, J.F., Gross, S., Melton, C., Fields, A.P., Liu, Q., et al. (2022). Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell* *40*, 1537–1549.e12.
63. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet.* *51*, 12–18.
64. Zviran, A., Schulman, R.C., Shah, M., Hill, S.T.K., Deochand, S., Khamnei, C.C., Maloney, D., Patel, K., Liao, W., Widman, A.J., et al. (2020). Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* *26*, 1114–1124.
65. Xiao, W., Ren, L., Chen, Z., Fang, L.T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., et al. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* *39*, 1141–1150.
66. Peneder, P., Stütz, A.M., Surdez, D., Krumbholz, M., Semper, S., Chircard, M., Sheffield, N.C., Pierron, G., Lapouble, E., Tötzel, M., et al. (2021). Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. *Nat. Commun.* *12*, 3230.
67. Moulriere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K., et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* *10*, eaat4921.
68. Esfahani, M.S., Hamilton, E.G., Mehrmohamadi, M., Nabet, B.Y., Alig, S.K., King, D.A., Steen, C.B., Macaulay, C.W., Schultz, A., Nesselbush,

- M.C., et al. (2022). Inferring gene expression from cell-free DNA fragmentation profiles. *Nat. Biotechnol.* **40**, 585–597.
69. Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., et al. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930.
70. Yousefi, P.D., Suderman, M., Langdon, R., Whitehurst, O., Davey Smith, G., and Relton, C.L. (2022). DNA methylation-based predictors of health: applications and statistical considerations. *Nat. Rev. Genet.* **23**, 369–383.
71. Maros, M.E., Capper, D., Jones, D.T.W., Hovestadt, V., von Deimling, A., Pfister, S.M., Benner, A., Zucknick, M., and Sill, M. (2020). Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat. Protoc.* **15**, 479–512.
72. Liu, M.C., Oxnard, G.R., Klein, E.A., Swanton, C., and Seiden, M.V.; CCGA Consortium (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759.
73. Nguyen, L., Van Hoeck, A., and Cuppen, E. (2022). Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat. Commun.* **13**, 4013.
74. Tang, W., Wan, S., Yang, Z., Teschendorff, A.E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* **34**, 398–406.
75. He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., Gao, W., Bing, P., Tian, G., and Yang, J. (2020). TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front. Bioeng. Biotechnol.* **8**, 394.
76. Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., PCAWG Tumor Subtypes and Clinical Translation Working Group, Danyi, A., De Ridder, J., van Herpen, C., Lolkema, M.P., and Steeghs, N. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728.
77. Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259.
78. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782.
79. Wang, K., Patkar, S., Lee, J.S., Gertz, E.M., Robinson, W., Schischlik, F., Crawford, D.R., Schäffer, A.A., and Ruppén, E. (2022). Deconvolving clinically relevant cellular immune cross-talk from bulk gene expression using CODEFACS and LIRICS stratifies patients with melanoma to anti-PD-1 therapy. *Cancer Discov.* **12**, 1088–1105.
80. Vorperian, S.K., Moufarrej, M.N., and Tabula Sapiens Consortium, and Quake, S.R. (2022). Cell types of origin of the cell-free transcriptome. *Nat. Biotechnol.* **40**, 855–861.
81. Chu, T., Wang, Z., Pe'er, D., and Danko, C.G. (2022 Apr). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat. Cancer* **3**, 505–517.
82. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474.
83. Klein, O., Kanter, F., Kulbe, H., Jank, P., Denkert, C., Nebrich, G., Schmitt, W.D., Wu, Z., Kunze, C.A., Sehoul, J., et al. (2019). MALDI-imaging for classification of epithelial ovarian cancer histotypes from a tissue microarray using machine learning methods. *Proteomics. Clin. Appl.* **13**, e1700181.
84. Sammut, S.J., Crispin-Ortuzar, M., Chin, S.F., Provenzano, E., Bardwell, H.A., Ma, W., Cope, W., Dariush, A., Dawson, S.J., Abraham, J.E., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629.
85. Kurtz, D.M., Esfahani, M.S., Scherer, F., Soo, J., Jin, M.C., Liu, C.L., Newman, A.M., Dührsen, U., Hüttmann, A., Casasnovas, O., et al. (2019). Dynamic risk profiling using serial tumor biomarkers for personalized outcome prediction. *Cell* **178**, 699–713.e19.
86. Nabet, B.Y., Esfahani, M.S., Moding, E.J., Hamilton, E.G., Chabon, J.J., Rizvi, H., Steen, C.B., Chaudhuri, A.A., Liu, C.L., Hui, A.B., et al. (2020). Noninvasive early identification of therapeutic benefit from immune checkpoint inhibition. *Cell* **183**, 363–376.e13.
87. Wu, Z., Lawrence, P.J., Ma, A., Zhu, J., Xu, D., and Ma, Q. (2020). Single-cell techniques and deep learning in predicting drug response. *Trends Pharmacol. Sci.* **41**, 1050–1065.
88. Wu, Z., Trevino, A.E., Wu, E., Swanson, K., Kim, H.J., D'Angio, H.B., Pre-ska, R., Charville, G.W., Dalerba, P.D., Egloff, A.M., et al. (2022). Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nat. Biomed. Eng.* **6**, 1435–1448.
89. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584.
90. Food and Drug Administration (2017). Discussion Paper on Laboratory Developed Tests (LDTs). <https://www.fda.gov/media/102367/download>.
91. Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., and Mann, R.M. (2019 Feb). Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* **290**, 305–314.
92. Baldwin, D.R., Gustafson, J., Pickup, L., Arteta, C., Novotny, P., Declerck, J., Kadir, T., Figueiras, C., Sterba, A., Exell, A., et al. (2020). External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax* **75**, 306–312.
93. Repici, A., Badalamenti, M., Maselli, R., Correale, L., Radaelli, F., Rondonotti, E., Ferrara, E., Spadaccini, M., Alkandari, A., Fugazza, A., et al. (2020). Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* **159**, 512–520.e7.
94. Lindpaintner, K., Mitchell, A., Pickering, C., Xu, G., Vignal, K., Axenfeld, B., Rice, R., Cong, X., Frederick, D.T., Michaud, W., et al. (2022). Glycoproteomics as a powerful liquid biopsy-based predictor of checkpoint inhibitor treatment benefit in metastatic malignant melanoma. *J. Clin. Orthod.* **40**, 9545.
95. Wu, Z., Serie, D., Xu, G., and Zou, J. (2020). PB-Net: Automatic peak integration by sequential deep learning for multiple reaction monitoring. *J. Proteomics* **223**, 103820.
96. Samir. Esophageal varices - post banding [Internet]. Wikimedia Commons. Available from: [https://commons.wikimedia.org/wiki/File:Esophageal\\_varices\\_-\\_post\\_banding.jpg](https://commons.wikimedia.org/wiki/File:Esophageal_varices_-_post_banding.jpg)
97. Henriksen, T.V., Tarazona, N., Frydendahl, A., Reinert, T., Gimeno-Vallente, F., Carbonell-Asins, J.A., Sharma, S., Renner, D., Hafez, D., Roda, D., et al. (2022). Circulating tumor DNA in stage III colorectal cancer, beyond minimal residual disease detection, toward assessment of adjuvant therapy efficacy and clinical behavior of recurrences. *Clin. Cancer Res.* **28**, 507–517.
98. Milbury, C.A., Creeden, J., Yip, W.K., Smith, D.L., Pattani, V., Maxwell, K., Sawchyn, B., Gjoerup, O., Meng, W., Skoletsky, J., et al. (2022). Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. *PLoS One* **17**, e0264138.
99. Dumur, C.I., Lyons-Weiler, M., Sciuilli, C., Garrett, C.T., Schrijver, I., Holley, T.K., Rodriguez-Paris, J., Pollack, J.R., Zehnder, J.L., Price, M., et al. (2008). Interlaboratory performance of a microarray-based gene

- expression test to determine tissue of origin in poorly differentiated and undifferentiated cancers. *J. Mol. Diagn.* 10, 67–77.
100. Homayounieh, F., Digumarthy, S., Ebrahimi, S., Rueckel, J., Hoppe, B.F., Sabel, B.O., Conjeti, S., Ridder, K., Siermanns, M., Wang, L., et al. (2021). An artificial intelligence-based chest X-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw. Open* 4, e2141096.
  101. Retson, T.A., Lim, V., and Watanabe, A.T. (2022). High performance of FDA-cleared platform for mammography triage. [https://hubspotusercontent20.net/hubfs/5209275/NCBC%202021%20cmTriage%20Poster%203\\_12\\_21\\_Final.pdf](https://hubspotusercontent20.net/hubfs/5209275/NCBC%202021%20cmTriage%20Poster%203_12_21_Final.pdf).
  102. Matthews, T.P., Singh, S., Mombourquette, B., Su, J., Shah, M.P., Pedemonte, S., Long, A., Maffit, D., Gurney, J., Hoil, R.M., et al. (2021). A multi-site study of a breast density deep learning model for full-field digital mammography and synthetic mammography. *Radiol. Artif. Intell.* 3, e200015.
  103. Wong, J., Fong, A., McVicar, N., Smith, S., Giambattista, J., Wells, D., Kolbeck, C., Giambattista, J., Gondara, L., and Alexander, A. (2020). Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiother. Oncol.* 144, 152–158.
  104. Liu, M.C., Oxnard, G.R., Klein, E.A., Swanton, C., and Seiden, M.V.; CCGA Consortium (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* 31, 745–759.
  105. Muehlematter, U.J., Daniore, P., and Vokinger, K.N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet. Digit. Health* 3, e195–e203.
  106. Center for Devices, Radiological Health (2022). Artificial Intelligence and Machine Learning Program: Research on AI/ML-Based Medical Devices (U.S. Food and Drug Administration, FDA). <https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/artificial-intelligence-and-machine-learning-program-research-aiml-based-medical-devices>.
  107. Lotter, W., Sorensen, G., and Cox, D. (2017). A multi-scale CNN and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer International Publishing), pp. 169–177.
  108. Pacilè, S., Lopez, J., Chone, P., Bertinotti, T., Grouin, J.M., and Fillard, P. (2020). Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol. Artif. Intell.* 2, e190208.
  109. Raciti, P., Sue, J., Ceballos, R., Godrich, R., Kunz, J.D., Kapur, S., Reuter, V., Grady, L., Kanan, C., Klimstra, D.S., and Fuchs, T.J. (2020). Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. *Mod. Pathol.* 33, 2058–2066.
  110. Pantanowitz, L., Quiroga-Garza, G.M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Albrecht Shach, A., Shalev, V., Vecsler, M., et al. (2020). An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *Lancet. Digit. Health* 2, e407–e416.
  111. Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S.E., Matin, R.N., Jain, A., Walter, F.M., Williams, H.C., and Deeks, J.J. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *BMJ* 368, m127.
  112. Archambault, Y., Boylan, C., Bullock, D., Morgas, T., Peltola, J., Ruokokoski, E., Genghi, A., Haas, B., Suhonen, P., and Thompson, S. (2020). Making on-line adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning. *Med. Phys. Intl. J.* 8.
  113. Bachar, N., Benbassat, D., Brailovsky, D., Eshel, Y., Glück, D., Levner, D., Levy, S., Pecker, S., Yurkovsky, E., Zait, A., et al. (2021). An artificial intelligence-assisted diagnostic platform for rapid near-patient hematology. *Am. J. Hematol.* 96, 1264–1274.
  114. Liu, X., Rivera, S.C., Moher, D., Calvert, M.J., and Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ* 370, m3164.
  115. Hwang, T.J., Sokolov, E., Franklin, J.M., and Kesselheim, A.S. (2016). Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. *BMJ* 353, i3323.
  116. Mishra, S. (2017). CE mark or something else?-Thinking fast and slow. *Indian Heart J. Teach. Ser.* 69, 1–5.
  117. Pesapane, F., Volontè, C., Codari, M., and Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* 9, 745–753.
  118. Oakden-Rayner, L., Gale, W., Bonham, T.A., Lungren, M.P., Carneiro, G., Bradley, A.P., and Palmer, L.J. (2022). Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet. Digit. Health* 4, e351–e358.
  119. Daneshjoo, R., He, B., Ouyang, D., and Zou, J.Y. (2021). How to evaluate deep learning for cancer diagnostics - factors and recommendations. *Biochim. Biophys. Acta. Rev. Cancer* 1875, 188515.
  120. Vodrahalli, K., Daneshjoo, R., Gerstenberg, T., and Zou, J. (2022). Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA (Association for Computing Machinery), pp. 763–777. (AI/ES '22).
  121. Ferryman, K. (2020). Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. *J. Am. Med. Inform. Assoc.* 27, 2016–2019.
  122. Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., and Miglioretti, D.L.; Breast Cancer Surveillance Consortium (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* 175, 1828–1837.
  123. Fenton, J.J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Intern. Med.* 175, 1837–1838.
  124. Duffy, G., Clarke, S.L., Christensen, M., He, B., Yuan, N., Cheng, S., and Ouyang, D. (2022). Confounders mediate AI prediction of demographics in medical imaging. *NPJ Digit. Med.* 5, 188.
  125. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215.
  126. Wulczyn, E., Steiner, D.F., Moran, M., Plass, M., Reihs, R., Tan, F., Flament-Auvigne, I., Brown, T., Regitnig, P., Chen, P.H.C., et al. (2021). Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit. Med.* 4, 71.
  127. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al. (2019). Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* 1, 236–245.
  128. Castelvocchi, D. (2016). Can we open the black box of AI? *Nature* 538, 20–23.
  129. Varghese, J. (2020). Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visc. Med.* 36, 443–449.
  130. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17, 195.
  131. Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G.P. (2018). Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* 174, 968–981.e15.
  132. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al.

- (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82.
133. Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwing, Y., Geller, L.T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., et al. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* 368, 973–980.
134. Beshnova, D., Ye, J., Onabolu, O., Moon, B., Zheng, W., Fu, Y.X., Brugarolas, J., Lea, J., and Li, B. (2020). De novo prediction of cancer-associated T cell receptors for noninvasive cancer detection. *Sci. Transl. Med.* 12, eaaz3738.
135. Levy-Jurgenson, A., Tekpli, X., Kristensen, V.N., and Yakhini, Z. (2020). Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. *Sci. Rep.* 10, 18802.
136. Vanguri, R.S., Luo, J., Aukerman, A.T., Egger, J.V., Fong, C.J., Horvat, N., Pagano, A., Araujo-Filho, J.d.A.B., Geneslaw, L., Rizvi, H., et al. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* 3, 1151–1164.
137. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *Adv. Neural Inf. Process. Syst.*, 5998–6008.
138. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., and Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 16144–16155.
139. Xiao, Y., Wu, J., and Lin, Z. (2021). Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Comput. Biol. Med.* 135, 104540.
140. Li, C.Y., Liang, X., Hu, Z., and Xing, E.P. (2019). Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *AAAI* 33, 6666–6673.

# Med

**Changing medicine.  
Together.**

*Med*, a new journal from Cell Press, publishes transformative, evidence-based science across the clinical and translational research continuum – from large-scale clinical trials to translational studies with demonstrable functional impact, offering novel insights in disease understanding.

We aim to elevate the global standard of medical research by accelerating translation of bench research to the clinic, serving as a hub for engagement between all stakeholders, improving reproducibility, and changing medical practice.

**Elevate your research. Submit your paper today.**

[cell.com/med](http://cell.com/med)



## Article

# A hybrid deep forest-based method for predicting synergistic drug combinations

Lianlian Wu,<sup>1,2,5</sup> Jie Gao,<sup>4,5</sup> Yixin Zhang,<sup>2,5</sup> Binsheng Sui,<sup>3</sup> Yuqi Wen,<sup>2</sup> Qingqiang Wu,<sup>3</sup> Kunhong Liu,<sup>3,\*</sup> Song He,<sup>2,\*</sup> and Xiaochen Bo<sup>1,2,6,\*</sup>

<sup>1</sup>Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin 300072, China

<sup>2</sup>Department of Bioinformatics, Institute of Health Service and Transfusion Medicine, Beijing 100850, China

<sup>3</sup>School of Film, Xiamen University, Xiamen 361005, China

<sup>4</sup>Department of Epidemiology and Health Statistics, School of Public Health, Fujian Medical University, Fuzhou 350122, China

<sup>5</sup>These authors contributed equally

<sup>6</sup>Lead contact

\*Correspondence: lkhqz@xmu.edu.cn (K.L.), hes1224@163.com (S.H.), boxc@bmi.ac.cn (X.B.)

<https://doi.org/10.1016/j.crmeth.2023.100411>

**MOTIVATION** Combination therapy has shown promise as a treatment for complex diseases such as cancer. Synergistic drug combinations can offer increased therapeutic efficacy and reduce toxicity compared with single drugs. However, class imbalances in datasets have complicated the use of computational tools, such as deep learning, for synergistic drug prediction. We propose an improved deep forest-based model, ForSyn, to address the above problem on imbalanced, medium- or small-scale datasets with high dimensionality.

## SUMMARY

Combination therapy is a promising approach in treating multiple complex diseases. However, the large search space of available drug combinations exacerbates challenge for experimental screening. To predict synergistic drug combinations in different cancer cell lines, we propose an improved deep forest-based method, ForSyn, and design two forest types embedded in ForSyn. ForSyn handles imbalanced and high-dimensional data in medium-/small-scale datasets, which are inherent characteristics of drug combination datasets. Compared with 12 state-of-the-art methods, ForSyn ranks first on four metrics for eight datasets with different feature combinations. We conduct a systematic analysis to identify the most appropriate configuration parameters. We validate the predictive value of ForSyn with cell-based experiments on several previously unexplored drug combinations. Finally, a systematic analysis of feature importance is performed on the top contributing features extracted by ForSyn. The resulting key genes may play key roles on corresponding cancers.

## INTRODUCTION

There has been important progress in anticancer drugs, especially targeted therapies. However, many tumors inevitably become resistant to the single agents.<sup>1–3</sup> To overcome the limitations of monotherapy, combination therapy has been proposed as a new treatment approach. In combination therapy, multiple drugs can target multiple targets, subpopulations, or diseases simultaneously.<sup>4,5</sup> Compared with monotherapy, combination therapy can increase therapeutic efficacy, reduce toxic side effects, and slow down the development of drug resistance.<sup>1,6–9</sup> For these therapeutic benefits, combination therapy has become a standard clinical treatment strategy for several complex diseases including cancers.<sup>7</sup>

Systematic surveys of effective drug combinations *in vitro* have been proposed such as the high-throughput screening

method.<sup>10</sup> However, it is insufficient for the large-scale experiments to search across such a large drug combination space.<sup>11–13</sup> To solve these problems, some computational approaches have been proposed such as network analysis<sup>14–16</sup> and mathematical models.<sup>17</sup> But most of them are often limited in the prior knowledge of biomedicine and the complexity of networks.<sup>14</sup> Alternatively, deep learning, as a data-driven computing method, has been widely used in drug combination prediction because of its generality, generalization and high prediction performance. Almost all deep learning methods used in drug combination prediction are based on deep neural networks (DNNs), including feedforward neural network,<sup>18,19</sup> deep belief network,<sup>20</sup> autoencoder,<sup>21</sup> transformer,<sup>22</sup> and graph neural network (GNN).<sup>23</sup> Although these methods have achieved high overall prediction performance, the problem of class imbalance is ignored. In drug combination dataset, the number of positive



samples (minority class) involving synergistic drug combinations is usually small. Although most samples are negative samples (majority class) including antagonistic, additive and slightly synergistic drug combinations, which is usually more than ten times the number of positive samples. Most previous methods are based on the assumption that the distribution of training samples in each class is balanced. In the case of imbalanced data, the classification results are usually biased toward the majority class.<sup>24,25</sup> That is, the model tends to predict more samples as majority (negative) class to obtain higher overall prediction accuracy, while ignoring the prediction accuracy on minority (positive) class. Especially in DNN-based methods, it is prone to overfitting because of the samples in minority class are particularly rare. Anand et al.<sup>26</sup> explored the impact of imbalanced data on the neural network backpropagation algorithms. They showed that the majority class essentially dominates the gradient of the network and is responsible for the weight update of the model. The classification error of the majority class will rapidly decrease in the early iteration process, while the classification error of the minority class will increase and cause the network to fall into a slow convergence mode.

In addition, most previous studies only applied structural and physicochemical properties of drugs, and gene expression profiles of untreated cancer cell lines to construct the feature set. This may ignore the biological connection between drugs and cancer cells, as synergism is the response of cells to drugs.<sup>5</sup> The response of cancer cells to drugs should also be considered.<sup>27–30</sup> Once more informative feature types are applied, the samples with missing features should be removed. The number of samples will be reduced, and the dimension of each sample's feature will be increased. The DNN-based methods always rely on the large-scale training datasets, and it is difficult to maintain its prediction performance on a medium- or small-scale dataset. Small sample size dataset with high dimensionality has further aggravated the difficulty in drug combination prediction. This is also an inherent problem in many biomedical datasets with multi-view/multi-omics data.

Given the powerful performance of deep learning technology on classification tasks, it is of great importance to explore the application of non-neural network deep learning technology on imbalanced, medium- or small-scale datasets with high dimensionality. Zhou et al.<sup>31</sup> proposed the deep forest (DF) model, which can be regarded as an alternative to DNN. DF is a multi-layer cascade structure, where each layer is composed of multiple tree-based forests. Each forest can be regarded as a unit in a cascade layer, similar to the neurons in the DNN. Compared with the DNN, the DF has the following advantages: suitable for datasets of different sizes, few hyper-parameters, and adaptive generation of model complexity.<sup>32</sup> The model complexity of DF can be adaptively determined under sufficient training. This advantage makes DF applicable to datasets of different scales, especially medium-sized datasets.<sup>33</sup> Because of its advantages, DF has been widely used in many fields, such as image retrieval,<sup>34</sup> cancer sub-category identification,<sup>35</sup> online financial cash-out monitoring,<sup>36</sup> etc. In the field of drug combination prediction, Zhang et al.<sup>37</sup> proposed a DF-based model, DCE-DForest, consisting of two components, a drug Bert<sup>38</sup> and a DF model. The Bert is a pretrained neural network to obtain the representations

of drugs, and a DF is used to predict drug combinations. First, the drug representations extracted by Bert cannot fully represent the multi-view (physical, biological, etc.) information of drugs. Each dimension of the representations has no specific meaning and cannot be interpreted. Second, DCE-DForest uses the original DF framework and does not consider the case of data with imbalance and high feature dimension.

To solve the above problems, we first construct a feature set consisting of physical, chemical and biological properties of drugs, in which the key features can be evaluated through ForSyn. The feature types include drug molecular fingerprints (DMFs), drug physicochemical properties (DPPs), cell line-specific drug-induced gene expression profiles (DGEs), and gene expression profiles of untreated cell lines (CGEs). The cell line-specific DGE feature can not only capture biological connection between drugs and cancer cells, but also be generalized to the study of patients.<sup>39</sup> Each dimension of the curated feature types has a specific meaning, which can facilitate the interpretable analysis to find out the key features in prediction process. Faced with this imbalanced, high-dimensional and medium-sized dataset, an improved DF-based model, ForSyn, is proposed to predict synergistic drug combinations. Two novel forest units are designed to embed in ForSyn. One is an RF based on affinity propagation (AP) clustering<sup>40</sup> and stratified under-sampling, which is designed to deal with the problem of class imbalance. The other is an extreme tree forest (ETF) that based on data complexity dimension reduction dealing with the problem of high-dimensional data. Then, the application of ForSyn is systematically analyzed by comparing 12 algorithms in eight datasets. The ForSyn with all the feature types wins the best performance in most cases. The performance of different configurations of ForSyn are also explored. Then, cellular experimental validation performed on a set of previously untested drug combinations further confirms the predictive ability of ForSyn. Finally, a systematic interpretable analysis of the key features extracted by ForSyn is performed.

## RESULTS

### The framework of ForSyn

In this study, the drug combinations tested in different cancer cell lines are collected as the sample dataset. The effects of drug combinations can be classified as synergism and non-synergism. A total of 3,192 samples are obtained from the DrugComb,<sup>41</sup> DrugCombDB,<sup>42</sup> and AstraZeneca-Sanger Drug Combination Prediction<sup>43</sup> databases, and classified according to the scheme proposed by Malyutina et al.<sup>44</sup> Two hundred samples are regarded as the synergism class (minority class), and the remaining 2,992 samples are classified as non-synergism class (majority class). The imbalance rate is close to 15, which is defined as the ratio between the size of the majority class and that of the minority class. Meanwhile, feature set is composed of four feature types. The 881-dimensional DMF, 55-dimensional DPP, and 978-dimensional DGE are used as the feature of drugs, the 978-dimensional CGE are used to represent cancer cell lines. All the feature types have been proved to be effective on other drug-related prediction tasks.<sup>33,45–49</sup> Each dimension of the curated feature types has a specific meaning,



**Table 1. Eight datasets used in this study**

Dataset	Description	Dimension
Data 1	DMF + CGE	2,740
Data 2	DPP + CGE	1,088
Data 3	DGE	1,956
Data 4	DMF + DPP + CGE	2,850
Data 5	DMF + DGE	3,718
Data 6	DPP + DGE	2,066
Data 7	DMF + DPP + DGE	3,828
Data 8	DMF + DPP + DGE + CGE	4,806

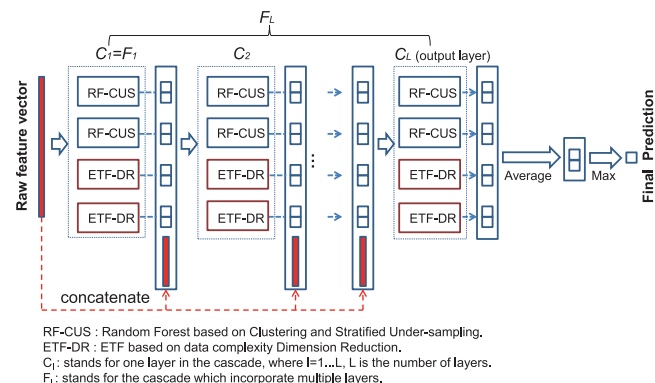
In the eight datasets, the representation of each sample is the concatenation of pairwise drug feature and the cell line feature.

which can facilitate the interpretable analysis to find out the key features in prediction process. More specifically, each dimension in DMF, DPP, and DGE represents a substructure, physico-chemical property, and gene expression values of drugs respectively.

To further investigate the influence of different representations in the classification process, eight different datasets including different feature combinations are generated (Table 1). In the training dataset of this study, a sample represent a drug combination on a particular cancer cell line (i.e., a drug combination-cell line pair). The same drug combination on different cell lines will have different effects. It is important to distinguish the drug combinations on different cell lines. In order to make the model to gain the distinguished ability, the representation of each sample is consisting of the drug feature and a cell line-specific feature. The drug feature includes DGE, DMF, and DPP. The cell line-specific features are DGE and CGE. According to the principle, eight datasets are generated and listed in Table 1.

Faced with the imbalanced, high-dimensional and medium-sized datasets, we propose ForSyn, which is a multi-layer cascade structure (Figure 1). Two novel forest types are embedded as the unit in each cascade layer. One is the RF based on clustering and stratified under-sampling (RF-CSU) dealing with imbalanced data. The other is an ETF based on data complexity dimension reduction (ETF-DR) dealing with high feature dimension (details are provided in STAR Methods).

RF is one of the representative algorithms of ensemble learning. It performs bootstrap sampling and random feature selection in the induction process of the base classifier. The perturbation of the feature space and the sample space ensures the diversity of the ensemble system. However, as with most traditional machine learning algorithms, the RF cannot effectively process imbalanced data. To deal with the problem of imbalanced data, the most common method is to rebalance the training set, such as randomly under-sampling the majority class. But this method always loses useful information. Some training samples that may play a key role in the classification process may be lost in the under-sampling process. To overcome this defect, we design an under-sampling method on the basis of AP clustering and stratified under-sampling, to rebalance the training set and minimize the information loss caused by random sampling. The proposed under-sampling method is



**Figure 1. The overall framework of ForSyn**

combined with the standard RF framework to rebalance the training set of each decision tree.

The ETF can be regarded as a variant of RF. Different from the RF, the ETF uses all the features as candidates, and then randomly selects a feature as the split node of the tree. The tree will continuously grow until each leaf node contains samples of the same class.<sup>32</sup> According to the properties, the ETF performs better to the imbalanced data. The pure leaf node that stores minority samples can effectively identify unknown minority samples. However, the high feature dimension and random selection of features would deepen the depth of the tree and cause over-fitting. To overcome the problem of ETF, we propose a greedy dimension reduction method, which combines a data complexity metric with the greedy algorithm. Data complexity, such as the shape of the decision boundary and the overlap between classes, is always used to describe the characteristics of the data.<sup>50</sup> The data complexity metrics would closely affect the predictive performance of the classifier.<sup>51</sup> In this study, the data complexity metric is defined as the tail overlap of the conditional distribution between two classes<sup>50</sup> (details are provided in STAR Methods).

### Performance evaluation

In this experiment, ForSyn is compared with 12 advanced algorithms on five metrics. The comparison algorithms include eight state-of-the-art deep learning-based algorithms in drug combination prediction, and four advanced machine learning algorithms. The deep learning-based algorithms are four DNN-based methods (DeepSynergy,<sup>18</sup> MatchMaker,<sup>19</sup> TranSynergy,<sup>22</sup> and SynPathy<sup>52</sup>), two DF-based methods (original DF<sup>32</sup> and DCE-DForest<sup>37</sup>), and two GNN-based methods (DeepDDS-GCN and DeepDDS-GAT<sup>23</sup>). The machine learning algorithms are two ensemble learning methods (XGBoost<sup>53</sup> and RF), and two imbalance learning methods (RUSBoost<sup>54</sup> and balanced bagging<sup>55</sup>). The evaluation metrics include F1 score, AUPR (area under the precision-recall curve), recall, MCC (Matthews correlation coefficient), and G-mean<sup>24</sup>; the F1 value is regarded as the main evaluation metric.

The results of all algorithms on the five metrics are shown in Tables S1–S5. The performance results are the mean value of ten-time 5-fold cross-validation (CV). In addition to the

DeepDDS, other 11 algorithms are tested on eight datasets (data 1–8) composed of different feature types. The performance of DeepDDS-GCN and DeepDDS-GAT based on graph data is shown in separate rows in Tables S1–S5. In Tables S1–S5, the performance of 11 algorithms based on data 1–8 is ranked, and the ranking values are shown in parentheses. The smaller ranking value indicates better performance. Then, the Friedman test and the Nemenyi test<sup>56</sup> are used to analyze the performance difference among the 11 algorithms. The Friedman test compares the performance differences of multiple algorithms on multiple datasets, while the Nemenyi test is performed between pairwise algorithms. According to Equations 15 and 16 (provided in STAR Methods), the Friedman statistical values and the corresponding *p* values in Tables S1–S5 are 16.40 ( $p = 2.638 \times 10^{-8}$ ), 55.35 ( $p = 5.200 \times 10^{-11}$ ), 37.30 ( $p = 1.480 \times 10^{-10}$ ), 28.10 ( $p = 1.823 \times 10^{-9}$ ), and 33.34 ( $p = 3.419 \times 10^{-10}$ ), respectively ( $N = 8, K = 11$ ). The distribution of  $F_F$  is based on the *F* distribution with 10 and 70 degrees of freedom. The critical value of  $F_F$  is 1.969 (Equation 16) with a 95% confidence level. The statistical results and *p* values on all the metrics reflect that there is a significant performance difference among the 11 algorithms. Next, according to Equation 17 and Table S6,  $CD = 5.338$  is calculated with the 95% confidence level in this study. Figures 2A–2E visually show the Nemenyi test results for Tables S1–S5. The average rank of the algorithms in data 1–8 is shown as the red dot in Figures 2A–2E.

From the average rank of performance results on data 1–8 (Figures 2A–2E; Tables S1–S5), it is observed that the ForSyn ranks first on four metrics, F1 score, AUPR, MCC and G-mean, showing its superior prediction performance. In addition, ForSyn performs better than the two DeepDDS algorithms on almost all datasets (Tables S1–S5). The deep learning-based algorithms, original DF, DCE-DForest, DeepSynergy, MatchMaker, TranSynergy, SynPathy and DeepDDS, have no module for imbalanced data processing, so the performance results on the five typical evaluation metrics of imbalanced data are relatively low. For the metric of recall, the performance of ForSyn ranks second, slightly lower than that of balanced bagging (Figure 2C; Table S3). Actually, the recall metric cannot fully reflect the performance of the model, and it often conflicts with precision. According to Figure 2A and Table S1, the F1 score of balanced bagging is low. It can be inferred that the algorithm greatly sacrifices the recognition rate of the majority class samples in exchange for an improvement in the recognition rate of the minority class samples. In addition to ForSyn, the other two DF-based algorithms, original DF and DCE-DForest, have similar ranks in all metrics and get the middle rank. This shows that the innovative design of ForSyn has brought great performance improvement. Figure 2F show the performance difference between ForSyn and other algorithms on the main metric (F1 score) more intuitively. From Figure 2F, it is observed that only three algorithms, XGBoost, random forest, and balanced bagging, have slightly better performance than ForSyn on data 1, 2 and 4. In addition to the three algorithms, ForSyn outperforms other comparison algorithms on all datasets. Similar results exist in other metrics. The performance difference on other metrics is shown in Figure S1.

Next, to evaluate the generalization performance on novel unseen cell lines, drugs and drug combinations, three cross-validation

strategies are performed. The training and test sets are shuffled by cell lines, drugs, or drug combinations, which are described as leave-cell-line-out CV, leave-drug-out CV and leave-drug-combination-out CV. The performance results are listed in Table S7. The result of the leave-drug-combination-out CV of all algorithms is inferior to random 5-fold cross-validation. For the leave-drug-out and leave-cell-line-out CV, the results are similar to those mentioned by Preuer et al.<sup>18</sup> That is, all methods yield low predictive performance and thus do not generalize well on novel drugs or novel cell lines, while ForSyn has achieved the best performance in F1 score, AUPR, and MCC, followed by TranSynergy. On the metric recall, RUSBoost is still the best, which is similar to the results discussed in random CV.

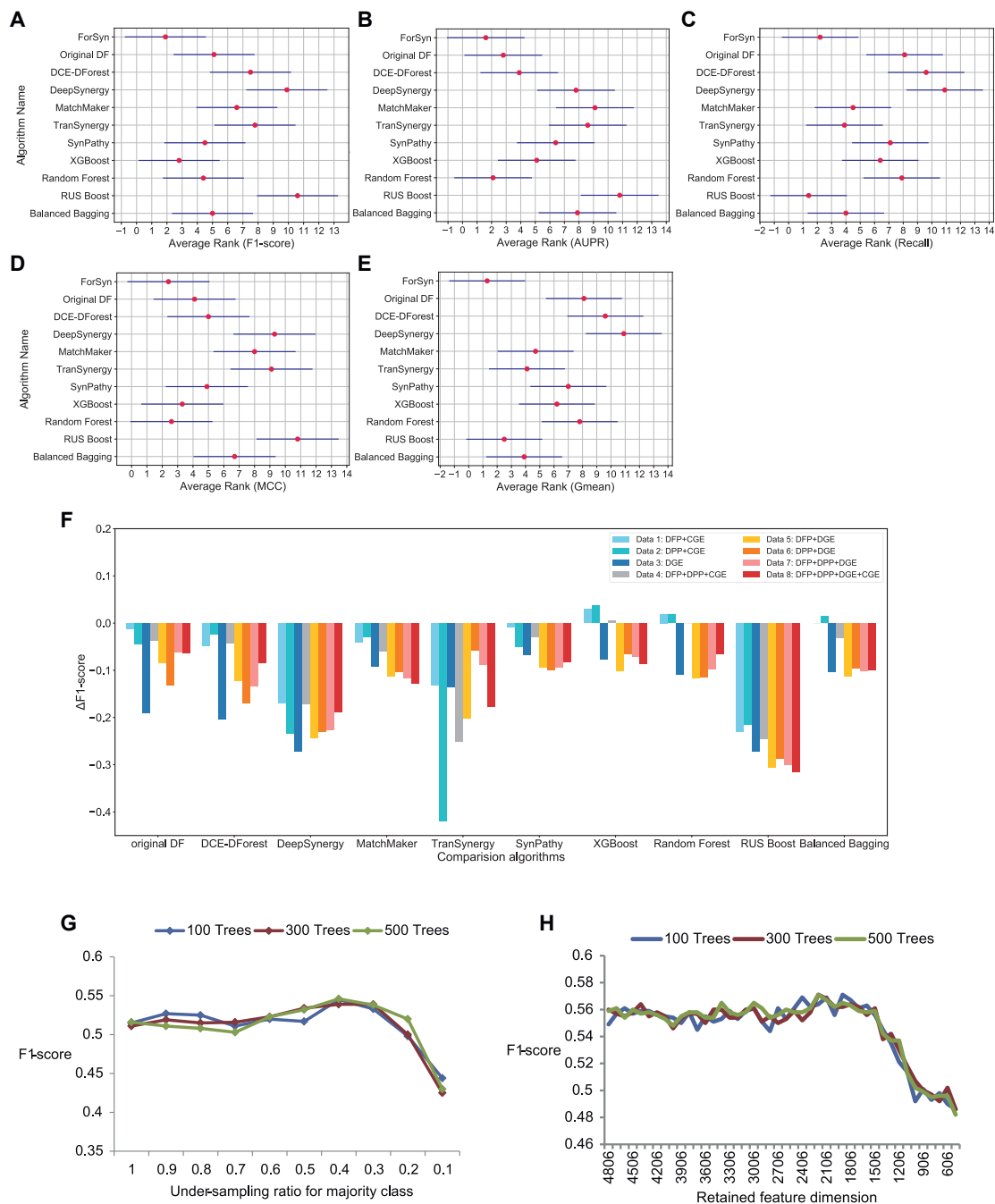
### Parameter analysis

Each layer of DF is the ensemble of multiple individual forests. In ForSyn, the RF-CSU unit dealing with data imbalance and the ETF-DR unit dealing with high-dimensional features are designed. This subsection will analyze the parameters that affect the performance of the RF-CUS, ETF-DR units, and ForSyn, respectively.

In the RF-CUS unit, the major parameter is the under-sampling ratio for the majority class, which is the ratio between the number of samples in the majority class before sampling and after sampling. We explore the effect of the number of base classifiers and under-sampling ratios on the performance of the RF-CSU unit. As shown in Figure 2G, the performance of the RF-CSU unit wins the best performance when the under-sampling ratio is 0.4. In addition, the increase in the number of decision trees does not bring a significant improvement in model performance. Therefore, in the ForSyn, the number of decision trees in the RF-CSU unit is set to 100, and the under-sampling ratio for the majority class is set to 0.4.

Figure 2H shows the parameters that affect the performance of the ETF-DR unit. The longest dimension of the samples is 4,806 (data 8). We first sort all features by data complexity, and then perform a greedy backward shrinkage to iteratively reduce the feature dimension, with a step size of 300. From Figure 2H, the performance of the ETF-DR unit wins the best performance when the retained dimension is 1,806. In addition, when the number of decision trees is set to 100, the model has the best performance. Therefore, in the ForSyn, we first reduce the feature dimension of the training sample to 1,806, then train the ETF-DR unit, and set the number of base classifiers of the ETF-DR unit to 100.

Table 2 shows the F1 score of the ForSyn under different configurations. It is observed that the average ranks of  $\text{ForSyn}^{(\text{RFC} \times 2 + \text{ETFD} \times 2)}$  and  $\text{ForSyn}^{(\text{RFC} \times 3 + \text{ETFD} \times 3)}$  are the same, and the average rank of  $\text{ForSyn}^{(\text{RFC} \times 4 + \text{ETFD} \times 4)}$  model is slightly lower. It is inferred that as the unit number increases, the performance of the model does not increase obviously. According to the principle of Occam's razor ("entities should not be multiplied unnecessarily"),  $\text{ForSyn}^{(\text{RFC} \times 2 + \text{ETFD} \times 2)}$  is chosen as the best configuration for the proposed model. In addition, by observing the performance of  $\text{ForSyn}^{(\text{RFC} \times 4)}$  and  $\text{ForSyn}^{(\text{ETFD} \times 4)}$ , it can be inferred that the ETF-DR unit has more advantages than the RF-CUS unit when processing drug combination dataset. If ForSyn



**Figure 2. Performance evaluation of ForSyn**

(A–E) According to Nemenyi test, the average rank of all algorithms tested on data 1–8 and five metrics: (A) F1 score, (B) AUPR, (C) recall, (D) MCC, and (E) G-mean. The average rank of each algorithm in eight datasets is marked as a red dot, and a horizontal line crossing the red dot indicates the range of CD value in Nemenyi test. The smaller the overlap between two horizontal bars, the more significant the difference between the two algorithms.

(F) The performance difference between ForSyn and other algorithms on F1 score under data 1–8. The y axis denotes  $\Delta F1$  between ForSyn and other comparison algorithms,  $\Delta F1 = F1_{\text{comparison algorithms}} - F1_{\text{ForSyn}}$ . A positive number indicates that the performance value of the comparison algorithm exceeds ForSyn, while a negative number indicates that ForSyn is superior to the comparison algorithm.

(G) The impact of the number of base classifiers and the under-sampling ratio on performance of ForSyn’s RF-CSU unit. The y axis represents the F1 score, and the x axis represents the under-sampling ratio for the majority class with a value range of 0.1–1. The blue, red, and green lines represent the RF-CUS unit containing 100, 300, and 500 decision trees, respectively.

(H) The impact of the number of base classifiers and the retained feature dimension on performance of ForSyn’s ETF-DR unit.

**Table 2. Performance of ForSyn in different configurations under F1 score**

	ForSyn <sup>(RFC*2+ETFD*2)</sup>	ForSyn <sup>(RFC*3+ETFD*3)</sup>	ForSyn <sup>(RFC*4+ETFD*4)</sup>	ForSyn <sup>(RFC*4)</sup>	ForSyn <sup>(ETFD*4)</sup>
Data 1	0.499 <sub>(2.5)</sub>	0.491 <sub>(4.0)</sub>	0.499 <sub>(2.5)</sub>	0.341 <sub>(5.0)</sub>	0.510 <sub>(1.0)</sub>
Data 2	0.496 <sub>(3.0)</sub>	0.525 <sub>(1.0)</sub>	0.501 <sub>(2.0)</sub>	0.364 <sub>(5.0)</sub>	0.477 <sub>(4.0)</sub>
Data 3	0.519 <sub>(3.0)</sub>	0.543 <sub>(1.5)</sub>	0.543 <sub>(1.5)</sub>	0.327 <sub>(5.0)</sub>	0.460 <sub>(4.0)</sub>
Data 4	0.529 <sub>(1.0)</sub>	0.524 <sub>(2.0)</sub>	0.519 <sub>(3.0)</sub>	0.349 <sub>(5.0)</sub>	0.475 <sub>(4.0)</sub>
Data 5	0.568 <sub>(3.0)</sub>	0.575 <sub>(1.5)</sub>	0.575 <sub>(1.5)</sub>	0.335 <sub>(5.0)</sub>	0.497 <sub>(4.0)</sub>
Data 6	0.551 <sub>(1.5)</sub>	0.539 <sub>(3.0)</sub>	0.551 <sub>(1.5)</sub>	0.345 <sub>(5.0)</sub>	0.473 <sub>(4.0)</sub>
Data 7	0.564 <sub>(1.5)</sub>	0.564 <sub>(1.5)</sub>	0.547 <sub>(3.0)</sub>	0.354 <sub>(5.0)</sub>	0.493 <sub>(4.0)</sub>
Data 8	0.572 <sub>(1.0)</sub>	0.556 <sub>(2.0)</sub>	0.547 <sub>(4.0)</sub>	0.339 <sub>(5.0)</sub>	0.551 <sub>(3.0)</sub>
Average rank	2.1	2.1	2.4	5.0	3.5

The value in parentheses represents the ranking value of the corresponding performance. Taking data 8 as an example, the ForSyn<sup>(RFC\*2+ETFD\*2)</sup> on this dataset has the best performance (0.572) and is assigned a ranking value of 1.0. In data 6, the performance of ForSyn<sup>(RFC\*2+ETFD\*2)</sup> and ForSyn<sup>(RFC\*4+ETFD\*4)</sup> are the same (0.551), and they occupy the first and second positions, respectively, so their ranking values are uniformly assigned 1.5 ((1.0 + 2.0)/2). The average rank of each algorithm is defined as the average of its ranks on all datasets. RFC, RF-CUS unit; ETFD, ETF-DR unit; and the number behind each unit represents the number of units of this type on each cascade layer. For example, ForSyn<sup>(RFC\*2+ETFD\*2)</sup> means that each cascade layer is placed with two RF-CUS units and two ETF-DR units.

only uses a single type of unit (for example, ForSyn<sup>(ETFD\*4)</sup>), it will cause the ensemble diversity of the cascade layer to decrease, which in turn leads to a decrease in the performance of the layer. However, the combination of different type units will promote the diversity of the cascade layer, which further improves the performance of the overall model.

Therefore, the optimal configuration of ForSyn is to place two RF-CUS units and two ETF-DR units in each cascade layer. Each RF-CUS unit contains 100 decision trees, and the under-sampling ratio is set to 0.4. Before training the ETF-DR unit, the proposed dimensionality reduction method is used to sort the feature space, and the first 1,806 dimensions of the sorted features are retained as the training set. The number of base classifiers in the ETF-DR unit is set to 100.

In addition, other tree-based forests are tested as the unit of ForSyn, including ADAboost (ADA), BAGging (BAG), and gradient boosting classifier (GBC). The base classifier of these models is the decision tree, and the parameters use default settings. Table 3 shows the performance comparison between the ForSyn and these derivative models. Under five evaluation metrics, the performance of the proposed ForSyn with two RF-CUS units and two ETF-DR units wins the best performance.

Subsequently, an ablation experiment on ForSyn is performed (Table S8). First, five different type units are placed on each cascade layer of DF, such as DF<sup>(ADA\*1+BAG\*1+GBC\*1+RF-CUS\*1+ETF-DR\*1)</sup>, the performance of this model can be regarded as a benchmark for ablation experiment (0.562). Then the units will be removed to observe the change of performance. As shown in Table S8, when the ETF-DR unit is removed, the model performance drops the most, followed by the RF-CUS unit. It can be inferred that the two units we designed are more suitable as units in the cascade framework than other decision tree ensembles.

### Cellular experiments of novel drug combinations

To confirm the efficacy of ForSyn, we further apply ForSyn to predict novel synergistic drug combination that have not been tested before. The cellular experiment is carried out on the pre-

dicted novel drug combinations. All drugs are combined in pairs, and the reported samples are removed. The remaining unmeasured samples are regarded as the novel drug combinations. According to the predicted probability of synergism class, eight drug combinations in the HT29 colorectal cell line with top predicted probability (Table S9) are selected to perform the cellular experiment. The synergistic potentials are observed on four drug combinations in the HT29 cell line, including erlotinib hydrochloride and AZD1775, erlotinib hydrochloride and MK-5108, etoposide and gefitinib, and erlotinib hydrochloride and dinaciclib (Figures 3A–3D).

Erlotinib hydrochloride is an inhibitor of the epidermal growth factor receptor tyrosine kinase (EGFR-TK). The EGFR has become an important therapeutic target for a variety of cancers.<sup>57,58</sup> The alterations of EGFR lead to cell growth, invasion, angiogenesis, and metastases. In colorectal cancer, 25%–77% of tumors overexpress EGFR.<sup>59,60</sup> There have been various EGFR inhibitors, such as erlotinib, an EGFR-TK inhibitor. Erlotinib has demonstrated efficacy against a range of solid tumor types including non-small-cell lung cancer (NSCLC), with more modest effects in colorectal cancer in phase I and II clinical trials.<sup>61–63</sup> Although the response rate of erlotinib is not satisfactory when used as monotherapy.<sup>64</sup> The combination therapy of erlotinib with other anticancer therapies should be more explored.

AZD1775 is a WEE1 inhibitor. It has been proved that the WEE1 gene could repair the DNA damage, which would limit the efficacy of DNA-damaging treatments in cancer cells.<sup>65</sup> The erlotinib has been found to suppress DNA damage repair in tumor cells.<sup>64</sup> The combination erlotinib and AZD1775 may enhance the sensitivity of tumor cells. MK-5108 is an Aurora-A kinase inhibitor. The synergistic effect has been observed in combined inhibition of the EGFR and Aurora-A pathways in cancer cells.<sup>66</sup> Aurora kinase inhibitors are active in combination with EGFR inhibition in a number of EGFR-mutant cell lines. Dinaciclib is a CDK inhibitor for CDK2, CDK5, CDK1, and CDK9. It has been reported that combined inhibition of EGFR and CDK9 resulted in reduced cell proliferation, accompanied by induction of apoptosis, G2-M cell-cycle arrest, inhibition of DNA

**Table 3. Performance comparison of deep forest embedding different units based on data 8**

Configuration	F1 score	AUPR	Recall	MCC	G-mean
DF <sup>(ADA*2+BAG*2)</sup>	0.500	0.582	0.384	0.509	0.614
DF <sup>(ADA*2+GBC*2)</sup>	0.484	0.559	0.350	0.508	0.588
DF <sup>(BAG*2+GBC*2)</sup>	0.493	0.561	0.365	0.512	0.598
DF <sup>(RF-CUS*2+ETF-DR*2)</sup>	0.572	0.591	0.537	0.535	0.722

DF, deep forest; ADA, ADAboost; BAG, BAGging; GBC, gradient boosting classifier.

replication and abrogation of CDK9-mediated transcriptional elongation, in contrast to monotherapy.<sup>67</sup>

In addition, giving gefitinib together with etoposide may kill more tumor cells (<https://clinicaltrials.gov>, NCT00483561). The phase II trial is studying how well giving gefitinib and etoposide works in treating patients with advanced prostate cancer that did not respond to hormone therapy. Gefitinib may stop the growth of tumor cells by blocking some of the enzymes needed for cell growth and by blocking blood flow to the tumor. Etoposide works in different ways to stop the growth of tumor cells, either by killing the cells or by stopping them from dividing.

Moreover, to investigate the potential false negatives of ForSyn, we also pick up five drug combinations in the HT29 cell line predicted as negative (non-synergistic) with highest probability to perform the same cellular experiment. The experimental results are shown in Figures 3E–3I. It is observed that all the five samples predicted by ForSyn as negative samples are verified as negative by cellular experiments. The CI of three drug combinations even exceeded 100, indicating the strong potential of non-synergistic. This further demonstrates the prediction accuracy of ForSyn in the negative samples.

### Interpretable analysis of feature importance

Model interpretation is of paramount importance in machine learning-based biomedical studies. In this study, ForSyn can evaluate the importance of each feature in the prediction process. ForSyn quantify the global relationship between each feature and the output by evaluating the feature importance value (FIV). Then, the FIVs extracted by ForSyn is analyzed from three aspects, the association with prediction process, the contribution of feature types, and the biological analysis of key features. All the FIVs are calculated on data 8 because it contains all the feature types.

#### Association with prediction process

First, two experiments are performed to show the relationship between FIVs and prediction process from the global and local perspectives, including the layer-by-layer error correction, and the difference of FIVs between different layers.

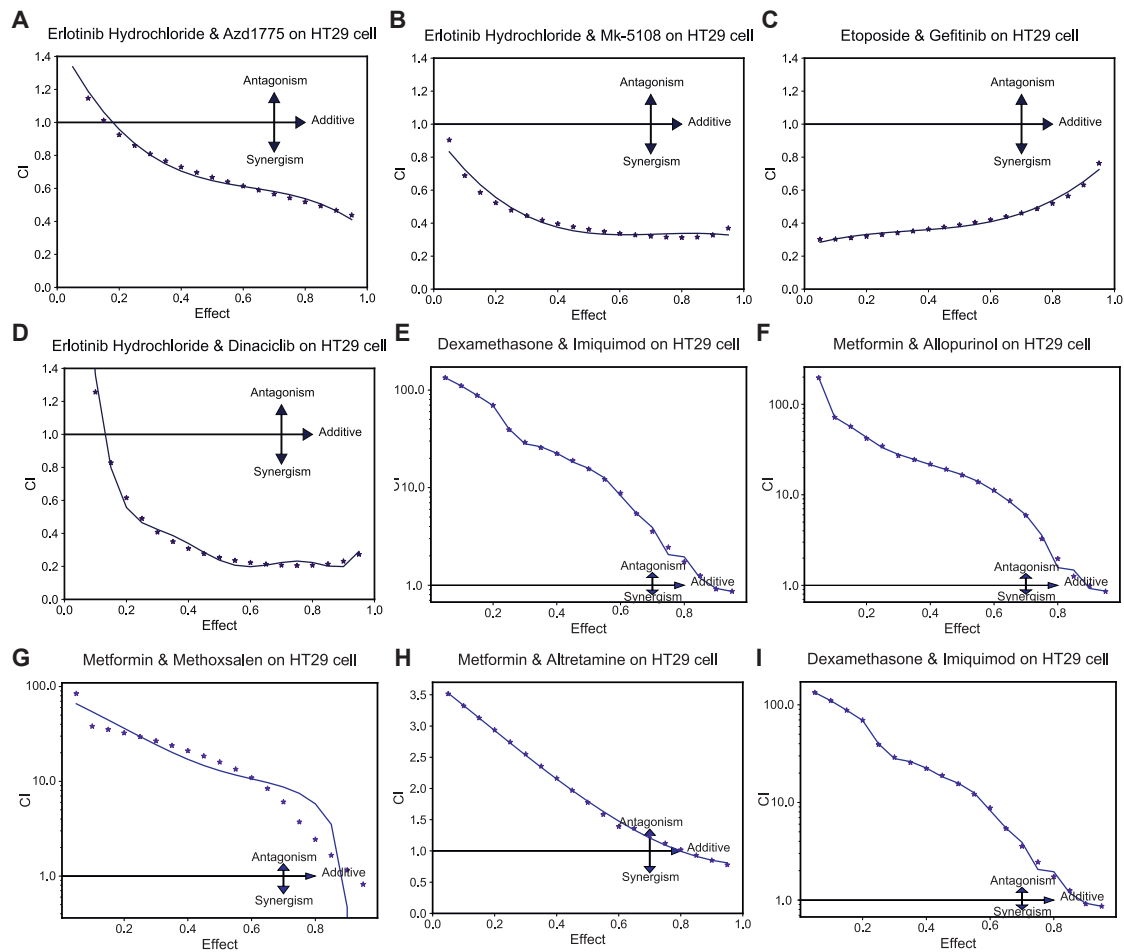
ForSyn is a deep learning method with multiple layers, which can be adaptively expanded according to the performance gain. In this section, we trained a ForSyn model with three layers, the classification results and FIVs of each layer in ForSyn are analyzed. In the first experiment, the layer-by-layer error correction capability of ForSyn is visualized in the feature space through FIVs (Figures 4A–4C). The positive samples (synergistic drug combinations) that are wrongly classified at each layer are

extracted. Then the top two features on the basis of the FIVs are used to project the mis-classified samples into a two-dimensional space. Figures 4A–4C shows the error correction result of each layer. The blue dots represent the mis-classified positive samples by the first layer of ForSyn. The red “+” represents the samples that are correctly classified at the second and last layers of ForSyn. From Figures 4A and 4B, the number of red plus signs appears more, indicating that the growth of the layer brings a significant performance improvement. In Figures 4B and 4C, the number of red plus signs increases slightly, indicating that the layer stops growing and the performance gradually converges. In addition, there are samples that cannot be corrected in the final layer, some of which may be related to the correctness of labels in the dataset. There may still be several incorrectly labeled noisy samples in the dataset because of the existence of experimental noise, as mentioned by Malyutina et al.<sup>68</sup>

In the local analysis of the association with prediction process, the difference of FIVs between different layers is evaluated. The FIV of each feature in the *l*th layer is calculated according to Equation 12 in STAR Methods. Then a rank vector is generated by sorting the FIVs of all features, so as to generate the rank vectors of three layers of ForSyn. Finally, the Wilcoxon signed rank test<sup>69</sup> is used to evaluate the significant differences between the three rank vectors. The *p* value for layer 1 vs. layer 2 is 0.958, and that for layer 2 vs. layer 3 is 0.972. The original hypothesis of this test is that there is no significant difference between paired vectors. Both *p* values are greater than 0.05, failing to reject the original hypothesis. That is, there is no significant difference between the paired FIVs’ rank vectors in the layers of ForSyn.

#### Contribution of feature types

The key features based on FIVs are then analyzed. The most contributing feature type is first investigated. The feature set is composed of four feature types, DMF, DPP, DGE, and CGE. When analyzing the FIVs, it should be noted that not all features participate in the whole prediction process. In the ETF-DR unit of ForSyn, a greedy dimension reduction method is applied to select 1,806-dimensional (see Parameter analysis) features to achieve the prediction task. Therefore, only the 1,806 features participate in the whole prediction process, including 1,037 DMFs, 31 DPPs, 600 DGEs, and 138 CGEs. The FIV of each feature is shown in Figure 4D. The red line in Figure 4D represents the average FIV of all features, which is 0.000554 (1/1,806). Figure 4E divides the features into two groups, the features that are greater than and less than the average FIV. It further shows the contribution of each feature type in the two groups. The contribution is calculated by summing the FIVs of features in a feature type. From Figures 4D and 4E, 768 features are greater than the average FIV, and the contributions of the 768 features are accounted for 74%. Therefore, we believe that these 768 features are top contributing features for prediction process. Among the 768 features, there are 107 DMFs, 17 DPPs, 582 DGEs, and 62 CGEs, with contributions of 15.35%, 2.86%, 49.70%, and 6.09%, respectively (Figure 4E). The results show that DGE plays a key role in the prediction process. Although there are many DMFs among 1,806 features, the contribution of most DMF features is lower than the average FIV (Figures 4D and 4E).



**Figure 3. The result of cellular experiment of ForSyn**

(A–D) The effect-CI plot of top predicted synergistic drug combinations tested in the HT29 cell line.  $CI < 1$  indicates that the drug combination has synergistic effect, while  $CI > 1$  indicates the non-synergistic effect.

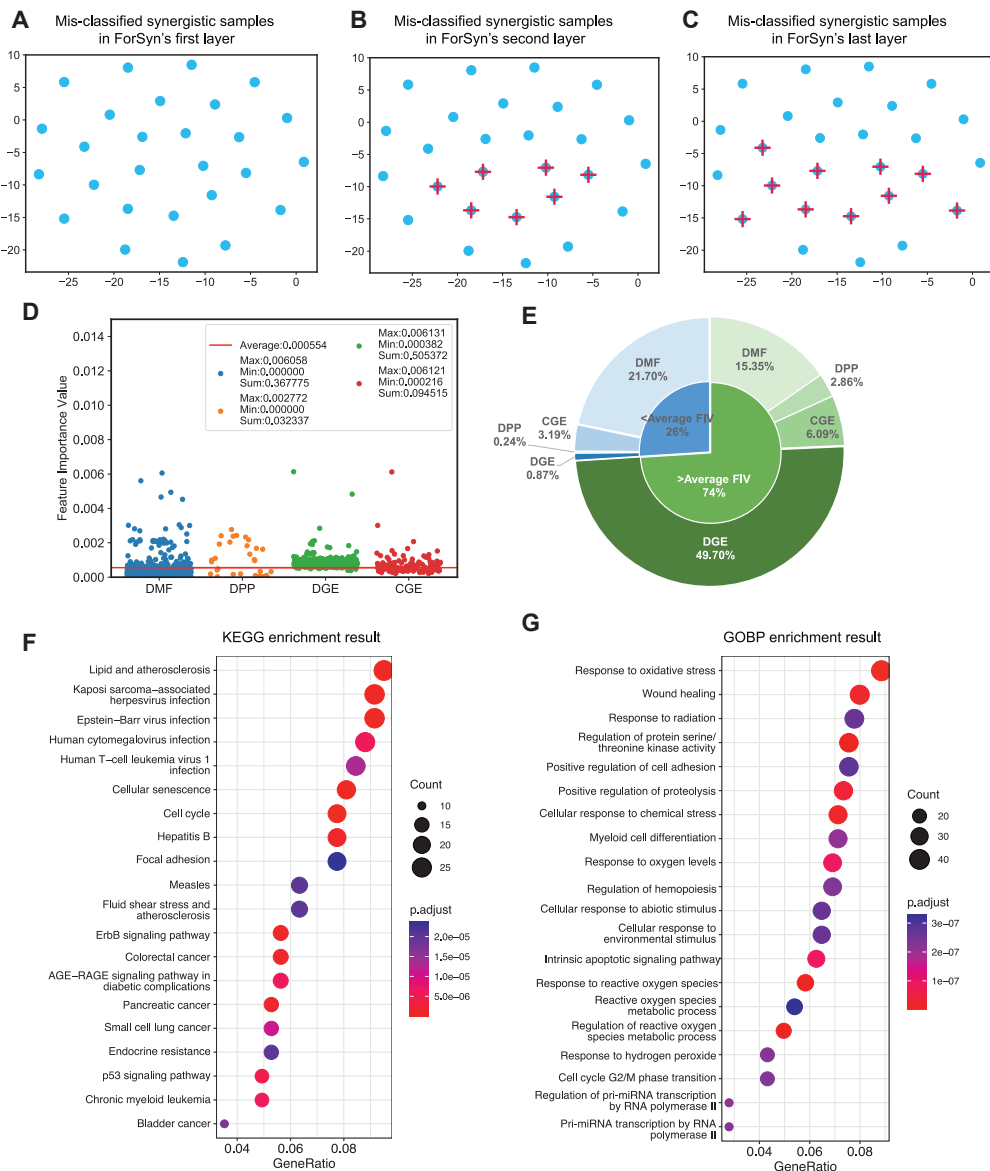
(E–I) The effect-CI plot of top predicted non-synergistic drug combinations tested in the HT29 cell line.

### Biological analysis of key features

Next, the biological analysis is performed on the key DGE features extracted by ForSyn. The 479 genes (with duplication removed) involved in the 582 DGE features that are greater than the average FIV are extracted. A global analysis on the extracted genes is carried out, including two kinds of gene enrichment analysis on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology Biological Process (GOBP). Enrichment results show that these genes are significantly enriched in 67 KEGG pathways and 518 GOBPs (adjusted  $p < 0.01$ ). The top 20 enrichment results are shown in Figures 4F and 4G. KEGG pathway enrichment result shows multiple significant biological pathways that are closely related to cancer (Figure 4F). According to the characteristics of these pathways, they can be divided into four categories: specific cancer pathways (colorectal cancer, pancreatic cancer, etc.), regulation process of cancer (cellular senescence, cell cycle, etc.), oncogenic virus infection (Kaposi sarcoma-associated herpesvirus infection,

etc.) and immune inflammation (lipid and atherosclerosis, etc.). For enrichment result of top 20 GOBP (Figure 4G), the key genes are more concentrated in the response to stimulus, especially the response to oxidative stress.

After the global analysis of the key genes, the cancer-specific key genes in DGE in different cell lines are further investigated. Four cell lines (HT29, A549, MCF7, and PC3) with more than 500 samples are selected to train the ForSyn respectively. Then the key DGE features with top FIVs of four cell lines are obtained. The top 10 genes involved in these key DGE features may play a key role in corresponding cancer cell lines, as shown in Table S10. For example, in A549 lung cancer cell line, CCND3 and TSPAN14 genes are identified as the top contributing genes. Song et al.<sup>70</sup> proposed that CCND3 could serve as potential biomarkers and provide a theoretical basis for the pathogenesis of lung adenocarcinoma. And TSPAN14 gene is also proposed as an indicator of NSCLC metastasis and progression.<sup>71</sup> In the HT29 colorectal cell line, the



**Figure 4. The interpretable analysis result of ForSyn**

(A–C) The top two features sorted by FIVs are used to visualize the ForSyn's layer-by-layer error correction of mis-classified positive (synergistic) samples. The blue dots represent the mis-classified positive samples by the first layer of ForSyn. The red plus sign represents the samples that are correctly classified at the second and last layers of ForSyn.

(D) The FIV of each feature in four feature types. The red line indicates the average FIV all features.

(E) The contribution of each feature type in two groups, which are the features that are greater than and less than the average FIV.

(F and G) The top 20 enrichment results of KEGG pathway and Gene Ontology Biological Process, which are obtained by key genes involved in the key DGE features.

CAMSAP2 gene has been proved to be a promising therapeutic target for the treatment of metastatic colorectal cancer patients.<sup>72</sup> PLOD3 has also been proved to be a potential biomarker for CRC diagnosis and prognosis prediction.<sup>73</sup> In MCF7 breast cancer cell and PC3 prostate carcinoma cell line, the top contributing genes, PGM1 and SPRED2, as well as SIRT3 and UFM1, are also proved to play a key role in breast and prostate cancers.<sup>74–78</sup>

## DISCUSSION

In this study, we propose a new algorithm, ForSyn, to predict synergistic drug combinations in different cancer cell lines. Two novel forest types are designed to embed in ForSyn, including the RF-CSU unit dealing with data imbalance and the ETF-DR unit dealing with high-dimensional features. The ForSyn can effectively solve the problems of class imbalanced,

and high feature dimension in the medium-scale datasets. Compared with 12 advanced algorithms on five metrics, ForSyn ranks first in four metrics, F1 score, AUPR, MCC and G-mean. Two statistical tests confirm that ForSyn perform significantly better than other algorithms in most cases. Next, the different configurations of ForSyn are analyzed. The results show that the under-sampling ratio for the majority class in RF-CSU, the feature dimension of the training sample in ETF-DR, the number of base classifiers, the types and numbers of units have influence on the performance of ForSyn. In addition, the novel synergistic drug combinations predicted by ForSyn are verified by cellular experiment, showing the predictive ability of ForSyn. Finally, a systematic interpretable analysis of the FIVs evaluated by ForSyn is performed. The layer-by-layer error correction and the difference of FIVs between different layers show the association between FIVs with prediction process. By summing the FIVs of each feature type, the DGE has been proved to play a critical role in the prediction process. Then the key genes involved in the key DGE features are explored by enrichment analysis. The key genes extracted by ForSyn may have potential effects on corresponding cancers.

Two forest types are designed in ForSyn, including RF-CSU and ETF-DR. The reason for choosing RF and ETF is that both models have their own advantages in dealing with high-dimensional and unbalanced data. The RF selects  $\sqrt{d}$  (where  $d$  is the dimension of the training data) features for each decision tree. Thus, the high feature dimension will not have a great negative impact on the performance of the RF, and effectively solving the problem of data imbalance is the key factor to improve the performance of RF. In the ETF model, the tree will continuously grow until each leaf node contains samples of the same class. Thus, the ETF has some advantages when dealing with imbalanced data. For example, the pure leaf node that stores minority samples can effectively identify unknown minority samples. However, the high feature dimension and the behavior of randomly selecting the feature, which deepens the depth of the tree and easily causes over-fitting. The effective dimension reduction methods may reduce computational cost and avoid over-fitting of the ETF. Thus, to obtain an excellent model to deal with the imbalanced and high-dimensional data, we design the modules of imbalanced data process and dimensionality reduction on RF and ETF respectively.

For the input feature data, the DGE and CGE can be quickly obtained at low cost through L1000 method or published predicted models when there are new drugs and cell lines to be predicted. In this study, the DGE and CGE are obtained from the National Institutes of Health (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS)<sup>79</sup> database. In LINCS database, the data are obtained using the L1000 method, which is a low-cost, high-throughput method and only needs 1,058 probes for 978 landmark transcripts and 80 control transcripts. The reagent cost of the L1000 assay is approximately \$2. The 978 landmarks have been shown to be sufficient to recover ~80% of the information in the full transcriptome. In addition, DGE and CGE also can be generated or predicted by machine learning models.<sup>80,81</sup> For example, Zhu et al.<sup>80</sup> have proposed a deep learning-based model, DLEPS, using SMILES of molecules to predict the 978-dimensional DGE obtained from LINCS database. DLEPS

has been validated in the use of screening potential drugs in obesity, hyperuricemia and nonalcoholic steatohepatitis.

ForSyn has shown an excellent predictive performance in drug combination prediction, which is validated by computational and biological experimental results. The novel units designed in ForSyn can largely solve the problems of imbalanced and high-dimensional data. Both are common problems in the datasets of drug-related biomedical studies. We hope that the propose of ForSyn can not only apply narrow down the candidates of drug combinations for experimental validations but also provide insights for other studies in drug discovery.

### Limitations of the study

Although ForSyn shows excellent prediction performance and interpretability, this study is limited by the number of training samples when using DGE and CGE as features. The importance of DGE has been shown in this study. In future work, we expect that the scale of the training dataset will expand with the accumulation of DGE, and the performance and interpretability of EC-DFR would be further improved accordingly. In addition, the predictive model cannot generalize well on novel drugs or novel cell lines, which is an inherent problem in drug combination prediction and should be explored in future work. Finally, some potential drug combinations and key genes has been found on the basis of ForSyn. The key factors should be further investigated through more biological experiments.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Drug combination dataset
  - Feature set
  - Description of ForSyn
  - Comparison methods
  - Evaluation metrics
  - Cross validations
  - Evaluation of feature importance value
  - Drug combination screening
- QUANTIFICATION AND STATISTICAL ANALYSIS

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.crmeth.2023.100411>.

### ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (grants 62103436 and 61772023), the National Key Research and Development Program of China (grant 2019QY1803), the Natural Science Foundation



of Fujian Province (grant 2022J01707), the High-Level Talents Research Start-Up Project of Fujian Medical University (grant XRCZX2021025), and the Fujian Science and Technology Plan Industry-University-Research Cooperation Project (grant 2021H6015).

### AUTHOR CONTRIBUTIONS

L.W., J.G., K.L., S.H., and X.B. designed the study and wrote the manuscript. L.W. and Y.W. acquired the data. J.G. and K.L. designed and applied the ForSyn model. L.W., J.G., B.S., and Q.W. analyzed the results. Y.Z. performed the cellular experiments. X.B., S.H., and K.L. supervised the research.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 1, 2022

Revised: November 27, 2022

Accepted: January 27, 2023

Published: February 21, 2023

### REFERENCES

- Jaaks, P., Coker, E.A., Vis, D.J., Edwards, O., Carpenter, E.F., Leto, S.M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., et al. (2022). Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* 603, 166–173. <https://doi.org/10.1038/s41586-022-04437-2>.
- Narayan, R.S., Molenaar, P., Teng, J., Cornelissen, F.M.G., Roelofs, I., Menezes, R., Dik, R., Lagerweij, T., Broersma, Y., Petersen, N., et al. (2020). A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nat. Commun.* 11, 2935. <https://doi.org/10.1038/s41467-020-16735-2>.
- Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., and Sarkar, S. (2014). Drug resistance in cancer: an overview. *Cancers* 6, 1769–1792. <https://doi.org/10.3390/cancers6031769>.
- Chou, T.-C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacol. Rev.* 58, 621–681. <https://doi.org/10.1124/pr.58.3.10>.
- Wu, L., Wen, Y., Leng, D., Zhang, Q., Dai, C., Wang, Z., Liu, Z., Yan, B., Zhang, Y., Wang, J., et al. (2022). Machine learning methods, databases and tools for drug combination prediction. *Briefings Bioinf.* 23, bbab355. <https://doi.org/10.1093/bib/bbab355>.
- Palmer, A.C., and Sorger, P.K. (2017). Combination cancer therapy can confer benefit via patient-to-patient variability without drug additivity or synergy. *Cell* 171, 1678–1691.e13. <https://doi.org/10.1016/j.cell.2017.11.009>.
- Ianevski, A., Giri, A.K., Gautam, P., Kononov, A., Potdar, S., Saarela, J., Wennerberg, K., and Aittokallio, T. (2019). Prediction of drug combination effects with a minimal set of experiments. *Nat. Mach. Intell.* 1, 568–577. <https://doi.org/10.1038/s42256-019-0122-4>.
- Sheng, Z., Sun, Y., Yin, Z., Tang, K., and Cao, Z. (2018). Advances in computational approaches in identifying synergistic drug combinations. *Briefings Bioinf.* 19, 1172–1182. <https://doi.org/10.1093/bib/bbx047>.
- Ramsay, R.R., Popovic-Nikolic, M.R., Nikolic, K., Uliassi, E., and Bolognesi, M.L. (2018). A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* 7, 3. <https://doi.org/10.1186/s40169-017-0181-2>.
- Lehár, J., Krueger, A.S., Avery, W., Heilbut, A.M., Johansen, L.M., Price, E.R., Rickles, R.J., Short, G.F., III, Staunton, J.E., Jin, X., et al. (2009). Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.* 27, 659–666. <https://doi.org/10.1038/nbt.1549>.
- Zhao, X.-M., Iskar, M., Zeller, G., Kuhn, M., van Noort, V., and Bork, P. (2011). Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Comput. Biol.* 7, e1002323. <https://doi.org/10.1371/journal.pcbi.1002323>.
- Bleicher, K.H., Böhm, H.J., Müller, K., and Alanine, A.I. (2003). Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* 2, 369–378. <https://doi.org/10.1038/nrd1086>.
- Morris, M.K., Clarke, D.C., Osimiri, L.C., and Lauffenburger, D.A. (2016). Systematic analysis of quantitative logic model ensembles predicts drug combination effects on cell signaling networks. *CPT Pharmacometrics Syst. Pharmacol.* 5, 544–553. <https://doi.org/10.1002/psp4.12104>.
- Feala, J.D., Cortes, J., Duxbury, P.M., Piermarocchi, C., McCulloch, A.D., and Paternostro, G. (2010). Systems approaches and algorithms for discovery of combinatorial therapies. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 181–193. <https://doi.org/10.1002/wsbm.51>.
- Cheng, F., Kovács, I.A., and Barabási, A.L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* 10, 1197. <https://doi.org/10.1038/s41467-019-09186-x>.
- Tang, J., Karhinen, L., Xu, T., Szwajda, A., Yadav, B., Wennerberg, K., Aittokallio, T., and Aittokallio, T. (2013). Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. *PLoS Comput. Biol.* 9, e1003226. <https://doi.org/10.1371/journal.pcbi.1003226>.
- Lee, J.-H., Kim, D.G., Bae, T.J., Rho, K., Kim, J.-T., Lee, J.-J., Jang, Y., Kim, B.C., Park, K.M., and Kim, S. (2012). CDA: combinatorial drug discovery using transcriptional response modules. *PLoS One* 7, e42573. <https://doi.org/10.1371/journal.pone.0042573>.
- Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics* 34, 1538–1546. <https://doi.org/10.1093/bioinformatics/btx806>.
- Kuru, H.I., Tastan, O., and Cicek, A.E. (2022). MatchMaker: a deep learning framework for drug synergy prediction. *IEEE ACM Trans. Comput. Biol. Bioinf* 19, 2334–2344. <https://doi.org/10.1109/TCBB.2021.3086702>.
- Chen, G., Tsoi, A., Xu, H., and Zheng, W.J. (2018). Predict effective drug combination by deep belief network and ontology fingerprints. *J. Biomed. Inf.* 85, 149–154. <https://doi.org/10.1016/j.jbi.2018.07.024>.
- Zhang, T., Zhang, L., Payne, P.R.O., and Li, F. (2021). Synergistic drug combination prediction by integrating multiomics data in deep learning models. *Methods Mol. Biol.* 2194, 223–238. [https://doi.org/10.1007/978-1-0716-0849-4\\_12](https://doi.org/10.1007/978-1-0716-0849-4_12).
- Liu, Q., and Xie, L. (2021). TranSynergy: mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations. *PLoS Comput. Biol.* 17, e1008653. <https://doi.org/10.1371/journal.pcbi.1008653>.
- Wang, J., Liu, X., Shen, S., Deng, L., and Liu, H. (2022). DeepDDS: deep graph neural network with attention mechanism to predict synergistic drug combinations. *Briefings Bioinf.* 23, bbab390. <https://doi.org/10.1093/bib/bbab390>.
- Johnson, J.M., and Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. *J. Big Data* 6, 27. <https://doi.org/10.1186/s40537-019-0192-5>.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* 73, 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>.
- Anand, R., Mehrotra, K.G., Mohan, C.K., and Ranka, S. (1993). An improved algorithm for neural network classification of imbalanced training sets. *IEEE Trans. Neural Network.* 4, 962–969. <https://doi.org/10.1109/72.286891>.
- Bollenbach, T., and Kishony, R. (2011). Resolution of gene regulatory conflicts caused by combinations of antibiotics. *Mol. Cell* 42, 413–425. <https://doi.org/10.1016/j.molcel.2011.04.016>.
- López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* 9, 583–593. <https://doi.org/10.1038/nrg2398>.

29. Geva-Zatorsky, N., Dekel, E., Cohen, A.A., Danon, T., Cohen, L., and Alon, U. (2010). Protein dynamics in drug combinations: a linear superposition of individual-drug responses. *Cell* **140**, 643–651. <https://doi.org/10.1016/j.cell.2010.02.011>.
30. Lukačičin, M., and Bollenbach, T. (2019). Emergent gene expression responses to drug combinations predict higher-order drug interactions. *Cell Syst.* **9**, 423–433.e3. <https://doi.org/10.1016/j.cels.2019.10.004>.
31. Zhou, Z.H., and Feng, J. (2017). Deep forest: towards an alternative to deep neural networks. *IJCAI*, 3553–3559.
32. Zhou, Z.-H., and Feng, J. (2019). Deep forest. *Natl. Sci. Rev.* **6**, 74–86. <https://doi.org/10.1093/nsr/nwy108>.
33. Lin, W., Wu, L., Zhang, Y., Wen, Y., Yan, B., Dai, C., Liu, K., He, S., and Bo, X. (2022). An enhanced cascade-based deep forest model for drug combination prediction. *Briefings Bioinf.* **23**, bbab562. <https://doi.org/10.1093/bib/bbab562>.
34. Zhou, M., Zeng, X., and Chen, A. (2019). Deep forest hashing for image retrieval. *Pattern Recognit. DAGM.* **95**, 114–127. <https://doi.org/10.1016/j.patcog.2019.06.005>.
35. Guo, Y., Liu, S., Li, Z., and Shang, X. (2017). Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data. *IEEE international conference on bioinformatics and biomedicine (BIBM)*, 1664–1669.
36. Zhang, Y.-L., Zhou, J., Zheng, W., Feng, J., Li, L., Liu, Z., Li, M., Zhang, Z., Chen, C., Li, X., et al. (2019). Distributed deep forest and its application to automatic detection of cash-out fraud. *ACM Trans. Intell. Syst. Technol.* **10**, 1–19. <https://doi.org/10.1145/3342241>.
37. Zhang, W., Xue, Z., Li, Z., and Yin, H. (2022). DCE-DForest: a deep forest model for the prediction of anticancer drug combination effects. *Comput. Math. Methods Med.* **2022**, 8693746. <https://doi.org/10.1155/2022/8693746>.
38. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: a robustly optimized bert pretraining approach. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1907.11692>.
39. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17. <https://doi.org/10.1016/j.cell.2017.10.049>.
40. Frey, B.J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* **315**, 972–976. <https://doi.org/10.1126/science.1136800>.
41. Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Maljutina, A., Jafari, M., Tanoli, Z., Pessia, A., and Tang, J. (2019). DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.* **47**, W43–W51. <https://doi.org/10.1093/nar/gkz337>.
42. Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.* **48**, D871–D881. <https://doi.org/10.1093/nar/gkz1007>.
43. Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* **10**, 2674. <https://doi.org/10.1038/s41467-019-09799-2>.
44. Maljutina, A., Majumder, M.M., Wang, W., Pessia, A., Heckman, C.A., and Tang, J. (2019). Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer. *PLoS Comput. Biol.* **15**, e1006752. <https://doi.org/10.1371/journal.pcbi.1006752>.
45. Zagidullin, B., Wang, Z., Guan, Y., Pitkänen, E., and Tang, J. (2021). Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings Bioinf.* **22**, bbab291. <https://doi.org/10.1093/bib/bbab291>.
46. Huang, Y.-A., You, Z.-H., and Chen, X. (2018). A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. *Curr. Protein Pept. Sci.* **19**, 468–478. <https://doi.org/10.2174/1389203718666161122103057>.
47. Hessler, G., and Baringhaus, K.-H. (2018). Artificial intelligence in drug design. *Molecules* **23**, 2520. <https://doi.org/10.3390/molecules23102520>.
48. Rifaioğlu, A.S., Cetin Atalay, R., Cansen Kahraman, D., Doğan, T., Martin, M., and Atalay, V. (2021). MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. *Bioinformatics* **37**, 693–704. <https://doi.org/10.1093/bioinformatics/btaa858>.
49. Xing, J., Shankar, R., Drelich, A., Paithankar, S., Chekalin, E., Dexheimer, T., Rajasekaran, S., Tseng, C.-T.K., and Chen, B. (2020). Reversal of infected host gene expression identifies repurposed drug candidates for COVID-19. Preprint at bioRxiv. <https://doi.org/10.1101/2020.04.07.030734>.
50. Cano, J.-R. (2013). Analysis of data complexity measures for classification. *Expert Syst. Appl.* **40**, 4820–4831. <https://doi.org/10.1016/j.eswa.2013.02.025>.
51. Sun, M., Liu, K., Wu, Q., Hong, Q., Wang, B., and Zhang, H. (2019). A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis. *Pattern Recognit. DAGM.* **90**, 346–362. <https://doi.org/10.1016/j.patcog.2019.01.047>.
52. Tang, Y.C., and Gottlieb, A. (2022). SynPathy: predicting drug synergy through drug-associated pathways using deep learning. *Mol. Cancer Res.* **20**, 762–769. <https://doi.org/10.1158/1541-7786.MCR-21-0735>.
53. Chen, T.Q., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
54. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., and Napolitano, A. (2010). RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. A.* **40**, 185–197. <https://doi.org/10.1109/TSMCA.2009.2029559>.
55. Hido, S., Kashima, H., and Takahashi, Y. (2009). Roughly balanced bagging for imbalanced data. *Stat. Anal. Data Min.* **2**, 412–426. <https://doi.org/10.1002/sam.10061>.
56. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30.
57. Meyerhardt, J.A., Zhu, A.X., Enzinger, P.C., Ryan, D.P., Clark, J.W., Kulke, M.H., Earle, C.C., Vincitore, M., Michelini, A., Sheehan, S., and Fuchs, C.S. (2006). Phase II study of capecitabine, oxaliplatin, and erlotinib in previously treated patients with metastatic colorectal cancer. *J. Clin. Oncol.* **24**, 1892–1897. <https://doi.org/10.1200/JCO.2005.05.3728>.
58. Mendelsohn, J., and Baselga, J. (2000). The EGF receptor family as targets for cancer therapy. *Oncogene* **19**, 6550–6565. <https://doi.org/10.1038/sj.onc.1204082>.
59. Mayer, A., Takimoto, M., Fritz, E., Schellander, G., Kofler, K., and Ludwig, H. (1993). The prognostic significance of proliferating cell nuclear antigen, epidermal growth factor receptor, and *mdr* gene expression in colorectal cancer. *Cancer* **71**, 2454–2460. [https://doi.org/10.1002/1097-0142\(19930415\)71:8<2454::AID-CNCR2820710805>3.0.CO;2-2](https://doi.org/10.1002/1097-0142(19930415)71:8<2454::AID-CNCR2820710805>3.0.CO;2-2).
60. Salomon, D.S., Brandt, R., Ciardiello, F., and Normanno, N. (1995). Epidermal growth factor-related peptides and their receptors in human malignancies. *Crit. Rev. Oncol. Hematol.* **19**, 183–232. [https://doi.org/10.1016/1040-8428\(94\)00144-1](https://doi.org/10.1016/1040-8428(94)00144-1).
61. Van Cutsem, E., Verslype, C., Beale, P., Clarke, S., Bugat, R., Rakhit, A., Fettner, S.H., Brennscheidt, U., Feyereislova, A., and Delord, J.P. (2008). A phase Ib dose-escalation study of erlotinib, capecitabine and oxaliplatin in metastatic colorectal cancer patients. *Ann. Oncol.* **19**, 332–339. <https://doi.org/10.1093/annonc/mdm452>.
62. Pérez-Soler, R., Chachoua, A., Hammond, L.A., Rowinsky, E.K., Huberman, M., Karp, D., Rigas, J., Clark, G.M., Santabarbara, P., and Bonomi,

- P. (2004). Determinants of tumor response and survival with erlotinib in patients with non-small-cell lung cancer. *J. Clin. Oncol.* *22*, 3238–3247. <https://doi.org/10.1200/JCO.2004.11.057>.
63. Tang, P.A., Tsao, M.-S., and Moore, M.J. (2006). A review of erlotinib and its clinical use. *Expert Opin. Pharmacother.* *7*, 177–193. <https://doi.org/10.1517/14656566.7.2.177>.
64. Zhang, Y., Zhou, F., Zhang, J., Zou, Q., Fan, Q., and Zhang, F. (2020). Erlotinib enhanced chemoradiotherapy sensitivity via inhibiting DNA damage repair in nasopharyngeal carcinoma CNE2 cells. *Ann. Palliat. Med.* *9*, 2559–2567. <https://doi.org/10.21037/apm-19-466>.
65. Watanabe, N., Broome, M., and Hunter, T. (1995). Regulation of the human WEE1Hu CDK tyrosine 15-kinase during the cell cycle. *EMBO J.* *14*, 1878–1891. <https://doi.org/10.1002/j.1460-2075.1995.tb07180.x>.
66. Niu, H., Manfredi, M., and Ecsedy, J.A. (2015). Scientific rationale supporting the clinical development strategy for the investigational Aurora A kinase inhibitor alisertib in cancer. *Front. Oncol.* *5*, 189. <https://doi.org/10.3389/fonc.2015.00189>.
67. McLaughlin, R.P., He, J., van der Noord, V.E., Redel, J., Foekens, J.A., Martens, J.W.M., Smid, M., Zhang, Y., and van de Water, B. (2019). A kinase inhibitor screen identifies a dual cdc7/CDK9 inhibitor to sensitize triple-negative breast cancer to EGFR-targeted therapy. *Breast Cancer Res.* *21*, 77. <https://doi.org/10.1186/s13058-019-1161-9>.
68. Malyutina, A., Majumder, M.M., Wang, W., Pessia, A., Heckman, C.A., and Tang, J. (2019). Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer. *PLoS Comput. Biol.* *15*, e1006752. <https://doi.org/10.1371/journal.pcbi.1006752>.
69. Taheri, S.M., and Hesamian, G. (2013). A generalization of the Wilcoxon signed-rank test and its applications. *Stat. Papers* *54*, 457–470. <https://doi.org/10.1007/s00362-012-0443-4>.
70. Song, Z., Zhang, Y., Chen, Z., and Zhang, B. (2021). Identification of key genes in lung adenocarcinoma based on a competing endogenous RNA network. *Oncol. Lett.* *21*, 60. <https://doi.org/10.3892/ol.2020.12322>.
71. Jovanović, M., Stanković, T., Stojković Burić, S., Banković, J., Dinić, J., Ljujić, M., Pešić, M., and Dragoj, M. (2022). Decreased TSPAN14 expression contributes to NSCLC progression. *Life* *12*, 1291. <https://doi.org/10.3390/life12091291>.
72. Wang, X., Liu, Y., Ding, Y., and Feng, G. (2022). CAMSAP2 promotes colorectal cancer cell migration and invasion through activation of JNK/c-Jun/MMP-1 signaling pathway. *Sci. Rep.* *12*, 16899. <https://doi.org/10.1038/s41598-022-21345-7>.
73. Shi, J., Bao, M., Wang, W., Wu, X., Li, Y., Zhao, C., and Liu, W. (2021). Integrated profiling identifies PLOD3 as a potential prognostic and immunotherapy relevant biomarker in colorectal cancer. *Front. Immunol.* *12*, 722807. <https://doi.org/10.3389/fimmu.2021.722807>.
74. Zheng, Z., Zhang, X., Bai, J., Long, L., Liu, D., and Zhou, Y. (2022). PGM1 suppresses colorectal cancer cell migration and invasion by regulating the PI3K/AKT pathway. *Cancer Cell Int.* *22*, 201. <https://doi.org/10.1186/s12935-022-02545-7>.
75. Vafeiadou, V., Hany, D., and Picard, D. (2022). Hyperactivation of MAPK induces tamoxifen resistance in SPRED2-deficient ERα-positive breast cancer. *Cancers* *14*, 954. <https://doi.org/10.3390/cancers14040954>.
76. Li, R., Quan, Y., and Xia, W. (2018). SIRT3 inhibits prostate cancer metastasis through regulation of FOXO3A by suppressing Wnt/beta-catenin pathway. *Exp. Cell Res.* *364*, 143–151. <https://doi.org/10.1016/j.yexcr.2018.01.036>.
77. Sawant Dessai, A., Dominguez, M.P., Chen, U.I., Hasper, J., Prechtel, C., Yu, C., Katsuta, E., Dai, T., Zhu, B., Jung, S.Y., et al. (2021). Transcriptional repression of SIRT3 potentiates mitochondrial aconitase activation to drive aggressive prostate cancer to the bone. *Cancer Res.* *81*, 50–63. <https://doi.org/10.1158/0008-5472.CAN-20-1708>.
78. Wei, Y., and Xu, X. (2016). UFMylation: a unique & fashionable modification for life. *Dev. Reprod. Biol.* *14*, 140–146. <https://doi.org/10.1016/j.gpb.2016.04.001>.
79. Pham, T.H., Qiu, Y., Zeng, J., Xie, L., and Zhang, P. (2021). A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat. Mach. Intell.* *3*, 247–257. <https://doi.org/10.1038/s42256-020-00285-9>.
80. Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., et al. (2021). Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat. Biotechnol.* *39*, 1444–1452. <https://doi.org/10.1038/s41587-021-00946-z>.
81. Stathias, V., Turner, J., Koleti, A., Vidovic, D., Cooper, D., Fazel-Najafabadi, M., Pilarczyk, M., Terryn, R., Chung, C., Umeano, A., et al. (2020). LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.* *48*, D431–D439. <https://doi.org/10.1093/nar/gkz1023>.
82. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* *171*, 1437–1452.e17. <https://doi.org/10.1016/j.cell.2017.10.049>.
83. Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., and Bryant, S.H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* *37*, W623–W633. <https://doi.org/10.1093/nar/gkp456>.
84. Cao, Y., Charisi, A., Cheng, L.C., Jiang, T., and Girke, T. (2008). ChemmineR: a compound mining framework for R. *Bioinformatics* *24*, 1733–1734. <https://doi.org/10.1093/bioinformatics/btn307>.
85. O’Boyle, N.M., Morley, C., and Hutchison, G.R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* *2*, 5. <https://doi.org/10.1186/1752-153x-2-5>.
86. Branco, P., Torgo, L., and Ribeiro, R.P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.* *49*, 1–50. <https://doi.org/10.1145/2907070>.
87. Davis, J., and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. <https://doi.org/10.1145/1143844.1143874>.
88. Al Iqbal, M.R., Rahman, S., Nabil, S.I., and Chowdhury, I.U.A. (2012). Knowledge based decision tree construction with feature importance domain knowledge. In *2012 7th International Conference on Electrical and Computer Engineering*, pp. 659–662. <https://doi.org/10.1109/ICECE.2012.6471636>.
89. Yuan, Y., Wu, L., and Zhang, X. (2021). Gini-Impurity index analysis. *IEEE Trans. Inf. Forensics Secur.* *16*, 3154–3169. <https://doi.org/10.1109/TIFS.2021.3076932>.

# Cell Reports Portfolio

Cutting-edge multidisciplinary  
research and methodologies

## Cell Reports Medicine

Translational and  
clinical research

## Cell Reports Physical Science

Research across the  
physical sciences

## Cell Reports

Research across  
the life sciences

## Cell Reports Methods

Methodological advances  
of broad interest

At Cell Press, we are committed to the principles of open research and to expanding our open access (OA) offering. That commitment began in 2012 when we launched *Cell Reports*, our first gold OA journal. Almost a decade later, the Cell Reports Portfolio has expanded to include journals that span life, medical, and physical science.



Explore more  
[cell.com/cell-reports-portfolio](https://cell.com/cell-reports-portfolio)



## Article

# Machine learning prediction of side effects for drugs in clinical trials

Diego Galeano<sup>1,4,\*</sup> and Alberto Paccanaro<sup>2,3</sup><sup>1</sup>Department of Electronics and Mechatronics Engineering, Facultad de Ingeniería, Universidad Nacional de Asunción, San Lorenzo, Paraguay<sup>2</sup>School of Applied Mathematics, Fundação Getulio Vargas, Rio de Janeiro, Brazil<sup>3</sup>Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham Hill, Egham, UK<sup>4</sup>Lead contact

\*Correspondence: dgaleano@ing.una.py

<https://doi.org/10.1016/j.crmeth.2022.100358>

**MOTIVATION** Drug side effects cause significant morbidity and mortality in healthcare. Side effects are discovered and added to the drug label during randomized controlled trials, but, due to trials' limited sample sizes, severe side effects are often discovered after the drug enters the market. An important question is whether we could use artificial intelligence to predict unknown side effects using the side effects identified during drug clinical trials. We studied this problem and developed a machine learning framework for predicting side effects for drugs undergoing clinical development.

## SUMMARY

Early and accurate detection of side effects is critical for the clinical success of drugs under development. Here, we aim to predict unknown side effects for drugs with a small number of side effects identified in randomized controlled clinical trials. Our machine learning framework, the geometric self-expressive model (GSEM), learns globally optimal self-representations for drugs and side effects from pharmacological graph networks. We show the usefulness of the GSEM on 505 therapeutically diverse drugs and 904 side effects from multiple human physiological systems. Here, we also show a data integration strategy that could be adopted to improve the ability of side effect prediction models to identify unknown side effects that might only appear after the drug enters the market.

## INTRODUCTION

Side effects of drugs are typically identified through randomized controlled clinical trials. It is well known that many side effects cannot be observed during clinical trials due to limitations in sample size and time frames. Postmarketing surveillance programs, such as the Adverse Event Reporting System (AERS), were designed to assist in the identification of side effects after the drug entered the market. However, the late identification of drug side effects is known to cause high morbidity and mortality in public healthcare,<sup>1,2</sup> the re-assessment of drug safety through new clinical trials,<sup>3</sup> and the possible withdrawal of drugs from the market.<sup>4</sup>

A wide range of computational approaches have been proposed to predict the side effects of drugs at different stages of the drug development process (see reviews by Ho et al.<sup>5</sup> and Boland et al.<sup>6</sup>). The first group of methods is applicable during pre-clinical drug development when only chemical, biological, and pharmacological information is available. These methods exploit chemical features,<sup>7–11</sup> protein targets,<sup>12</sup> and pathway

information,<sup>13</sup> often in combination with protein networks,<sup>14</sup> and, in general, they offer a modest accuracy. A second group of methods was proposed for the postmarketing phase of drug development.<sup>15–19</sup> These methods exploit the side effects collected in clinical trials and the postmarketing phase to predict other unknown side effects. Our study differs from these methods in that we assumed that only side effects identified during clinical trials are available. This represents a more challenging scenario due to information sparsity and selection bias.<sup>20,21</sup> Our goal is 2-fold: (1) to simulate the realistic scenarios faced by safety professionals working in clinical drug development and (2) to provide a computational tool that can assist in the early detection of side effects of drugs undergoing clinical trials.

A critical application of our approach is during the different phases of clinical trials, where computational predictions can be used as a hypotheses generator to set the direction of the risk assessment. Our approach uses a matrix completion model that we called the geometric self-expressive model (GSEM). This is based on our objective function and multiplicative learning algorithm, which learns globally optimal solutions. Our model



exploits known drug side effect associations and integrates graph structure information from chemical, biological, and pharmacological data. Here, we also show that predicting side effects that were identified after the drug entered the market from the information available during clinical trials is challenging. We attributed this to a distribution shift in side effect reports between clinical trials and postmarketing. This observation motivated a simple data integration technique that can be used to significantly improve the performance of GSEM at identifying side effects that might appear after the drug enters the market.

## RESULTS

### GSEM

Our starting point is the  $n \times m$  drug side effect association matrix  $X$ , where  $x_{ij} = 1$  if drug  $i$  is known to induce side effect  $j$ , or  $x_{ij} = 0$  otherwise. Drugs can be related by their similarities in chemical structure, biological targets, and pharmacological activity. Side effects can also be related by their similarities in anatomical/physiological phenotypes. Our method integrates drug and side effect information by learning two similarity matrices: a drug similarity matrix  $H \in \mathbb{R}^{n \times n}$  such that  $X \approx HX$  and a side effect similarity matrix  $W$  such that  $X \approx XW$ . The GSEM generates scores for each drug-side effect pair by linearly combining these models:

$$\hat{X} = HX + XW. \quad (\text{Equation 1})$$

The first term in Equation 1 is the drug self-representation model, and the second term is the side effect self-representation model. To learn  $W$  and  $H$ , we minimize the following objective functions:

$$\min_W \underbrace{\frac{1}{2} \|X - XW\|_F^2}_{\text{self-representation}} + \underbrace{\frac{a}{2} \|W\|_F^2 + b \|W\|_1}_{\text{sparsity}} + \underbrace{\sum_i \frac{\mu_i}{2} \|W\|_{D, \mathcal{G}_i}^2}_{\text{smoothness}} + \underbrace{\gamma \text{Tr}(W)}_{\text{diagonal}}$$

such that  $W \geq 0$

(Equation 2)

and

$$\min_H \underbrace{\frac{1}{2} \|X - HX\|_F^2}_{\text{self-representation}} + \underbrace{\frac{c}{2} \|H\|_F^2 + d \|H\|_1}_{\text{sparsity}} + \underbrace{\sum_j \frac{\alpha_j}{2} \|H\|_{D, \mathcal{G}_j}^2}_{\text{smoothness}} + \underbrace{\gamma \text{Tr}(H)}_{\text{null diagonal}}$$

such that  $H \geq 0$

(Equation 3)

where  $\|\cdot\|_F$  denotes the Frobenius norm. We shall explain each term in Equation 2 only, as the same rationale can be applied to Equation 3. The first term in Equation 2 is the self-representation constraint, which aims at learning a self-representation matrix  $W$  such that  $XW$  is a good reconstruction of the original matrix  $X$ . The second term, in which  $a, b > 0$  are constant values, is the sparsity constraint, which uses the elastic-net regularization known to impose sparsity and grouping effect.<sup>22,23</sup> The third term in Equation 2 is the smoothness constraint,<sup>24–26</sup> incorporating geometric structure into the self-representation matrix  $W$  from a given side effect similarity graph  $\mathcal{G}_i$ , with  $\mathcal{G}_i = (\{1, \dots,$

$m\}, \mathcal{E}_i, A_i)$ , i.e., the weighted undirected graph with edge weights  $A_{ij} > 0$  if  $(i, j) \in \mathcal{E}$  and zero otherwise. The smoothness constraint is important because it allows us to integrate into the model side information about side effects in the form of graphs. For a given side effect graph  $\mathcal{G}$ , the idea is that nearby points in  $\mathcal{G}$  should have similar coefficients in  $W$ , which can be obtained by minimizing

$$\sum_{ij} A_{ij} \|w_i - w_j\|^2 = \text{Tr}(WLW^T) := \|W\|_{D, \mathcal{G}}^2, \quad (\text{Equation 4})$$

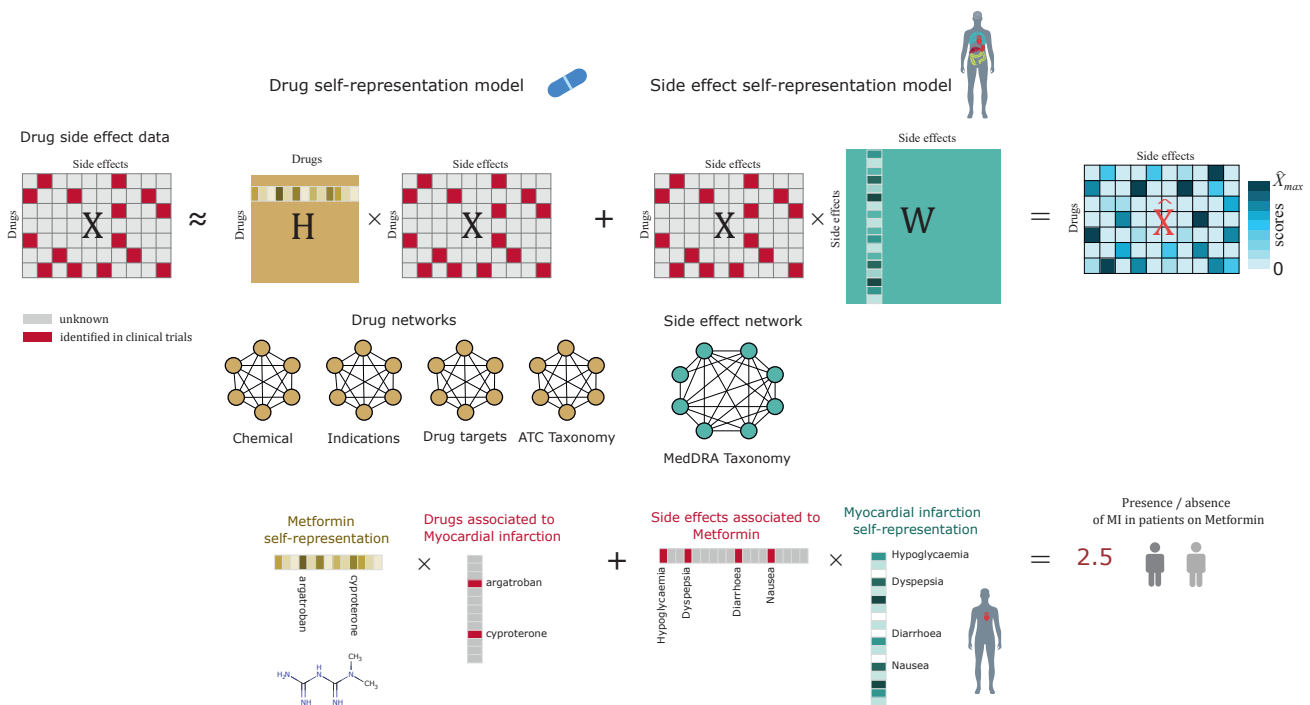
where  $w_i$  and  $w_j$  represent column vectors of  $W$  and  $L = D - A$  is the graph Laplacian with  $D = \text{diag}(\sum_j a_{ij})$ . The constant values  $\mu_i > 0$  in Equation 2 weigh the importance of the smoothness constraint for the prediction. When multiple graphs are combined, the parameters  $\mu_i$  in Equation 2 tell us about the contribution and importance of the individual graph information for the prediction model. The fourth term in Equation 2 is a penalty for diagonal elements to prevent the trivial solution  $W = I$  (the identity matrix). Typically,  $\gamma \gg 0$  is used. The last constraint in Equation 2 is a non-negative constraint,<sup>27</sup> which is added here to favor interpretability of the learned  $W$ .

Figure 1 depicts an overview of our GSEM. The starting point is the matrix  $X$  containing binary associations encoding the presence or absence of drug side effects. The GSEM learns the self-representation matrices  $H$  and  $W$  that minimize our loss functions in Equation 3 and 2, respectively, by employing an iterative algorithm that uses a simple multiplicative update rule (see STAR Methods). Our algorithm is inspired by the diagonally rescaled principle of non-negative matrix factorization.<sup>27</sup> GSEM is fast to run, and it does not require setting a learning rate or applying a projection function. Our algorithm also satisfies global guarantees of convergence given by the Karush-Kuhn-Tucker (KKT) complementary conditions (proof in Methods S2). Having learned independently  $H$  and  $W$ , we calculate  $\hat{X} = HX + XW$ . Notice that while  $X$  contains binary values  $[0, 1]$  that correspond to our original data,  $\hat{X}$  contains real positive numbers that are our predicted scores.

### Overview of evaluation

To obtain side effects identified in clinical trials, we followed the procedure in Galeano et al.<sup>28</sup> to retrieve side effects reported in randomized controlled studies from the Side Effect Resource (SIDER) 4.1.<sup>21</sup> 27,610 associations were obtained for  $n = 505$  marketed drugs and  $m = 904$  unique side effect terms. We also collected side effects identified after the drugs entered the market from two independent sources. 6,818 side effects reported in the postmarketing section of drug leaflets were obtained from the SIDER database (SIDER postmarket set). 25,797 statistically significant side effects reported in the AERS were obtained from the OFFSIDES database<sup>29</sup> (OFFSIDES postmarket set). The collection of drug side effect data used in our study is shown in Figure 2A.

Our goal is to assess the performance of the GSEM at predicting unknown side effects for drugs with a small number of side effects identified in clinical trials. Therefore, only side effects identified in clinical trials were used for training the model. Figure 2B illustrates how the clinical trials' side effects were randomly split into training, validation, and testing sets.



**Figure 1. Geometric self-expressive model (GSEM)**

27,610 associations identified on clinical trials for 505 drugs and 904 side effects were collected from the SIDER 4.1 database. The associations were arranged into an  $n \times m$  matrix  $X$  by encoding their presence ( $= 1$ ). Unknown associations were encoded with zeros ( $= 0$ ). Our algorithm learns two similarity matrices that model the two pharmacological spaces of drug side effects.  $H$  (of size  $n \times n$ ) encodes similarities between drugs that are learned from drug networks built from chemical, indication, target, and taxonomy similarities.  $W$  (of size  $m \times m$ ) encodes similarities between side effects that are learned from physiological relationships between side effects. The GSEM learns independently  $H$  and  $W$  such that  $X \approx HX$  and  $X \approx XW$ . By linearly combining these models,  $HX + XW$ , we obtain  $\hat{X}$ , which models  $X$ , and where all the entries are replaced by real numbers—these are our predicted scores. Note that values replacing zero entries in  $X$  will constitute our predictions. Rows of  $H$  are drug self-representations, and columns of  $W$  are side effect self-representations. The lower illustration depicts how our model discovers a drug self-representation vector for the anti-diabetic drug metformin, and a self-representation vector for the side effect myocardial infarction (MI), such that the dot product of these vectors with the binary vector corresponding to known drugs for MI and known side effects of metformin, respectively, models the presence/absence of MI in patients on metformin. The body parts infographic vector was created by macrovector [www.freepik.com](http://www.freepik.com).

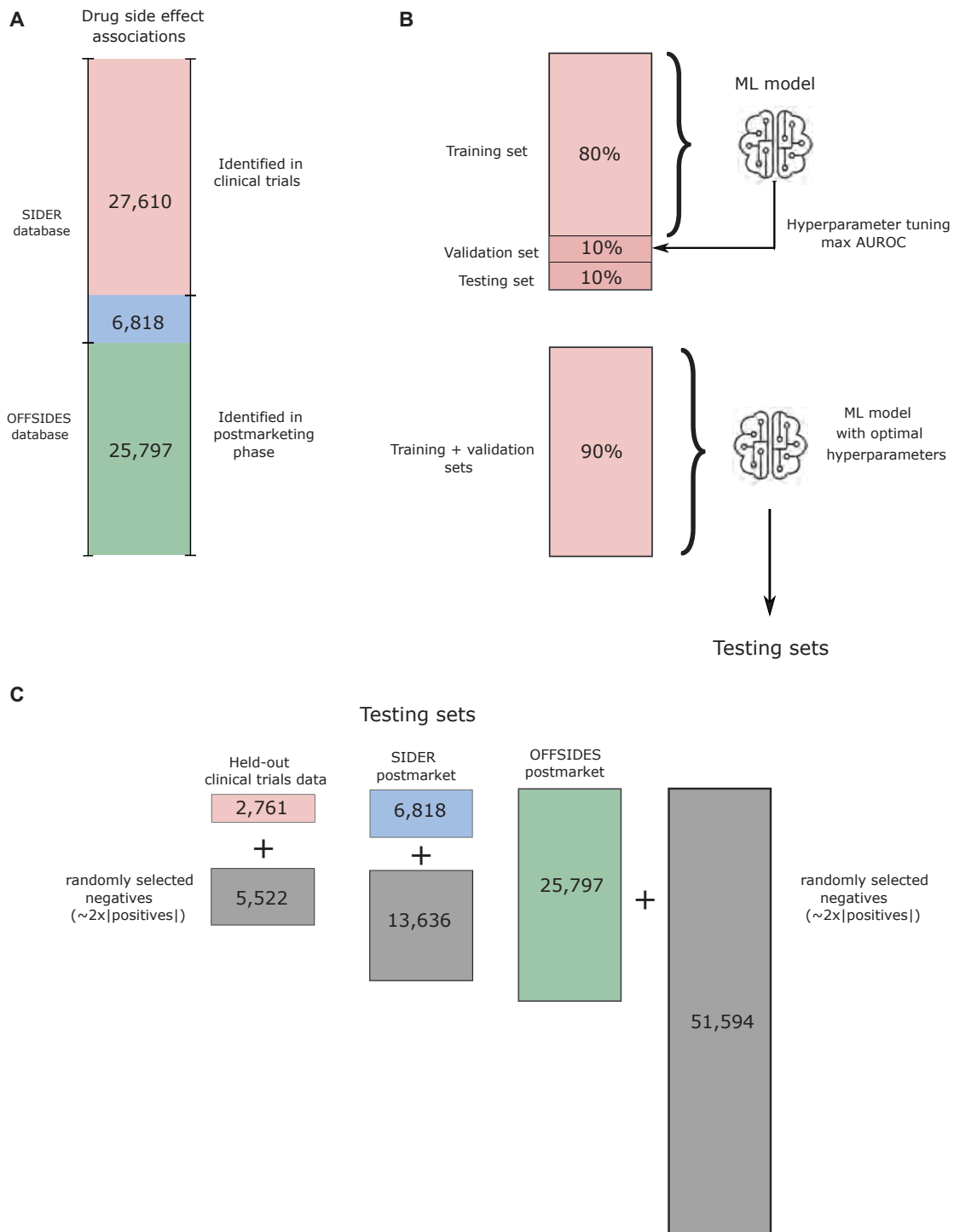
Following previous approaches,<sup>15–19</sup> we framed our problem as a binary classification problem and used the area under the receiving operating curve (AUROC). The validation set consisted of 10% randomly held-out clinical trials side effects and randomly selected negatives of twice the number of positives. We used the validation set to tune the model hyperparameters. We then performed the evaluation by training the model with the combined training and validation sets using the optimal hyperparameters. We measure the AUROC and the area under the precision-recall curve (AUPR) on three test sets (see Figure 2C): (1) a held-out test set from randomly selected side effects identified in clinical trials, (2) postmarketing side effects from the SIDER database, and (3) postmarketing side effects from the OFFSIDES database.

We compared the prediction performance of the GSEM with a representative number of side effect prediction models that can also be applied to our problem: (1) matrix factorization (MF);<sup>16</sup> (2) predictive pharmacosafety networks (PPNs);<sup>15</sup> (3) inductive matrix completion (IMC);<sup>17</sup> and (4) feature graph-regularized MF (FGRMF).<sup>18</sup> Each side effect prediction model integrates different types of complementary information about drugs and

side effects. We collected and used five types of side information for our study. For drugs, we obtained the chemical structure and protein targets from DrugBank,<sup>30</sup> indications from the Drug Repositioning Hub,<sup>31</sup> and Anatomical, Therapeutic, and Chemical (ATC) classification (see STAR Methods). We used MACCS fingerprints<sup>32</sup> to represent chemical structure and computed Tanimoto similarity using RDKit.<sup>33</sup> For side effects, we obtained the Medical Dictionary for Regulatory Activities (MedDRA) terminology. To build graphs from the different side information, we calculated the adjacency matrices using similarity measures (see STAR Methods). For the ATC and MedDRA terms, we also obtained their corresponding hierarchies to calculate taxonomy similarities that have been used by previous approaches.<sup>15,17</sup>

### Evaluation of prediction performance on multiple drugs

Figure 3A shows the AUROC performance of the side effect prediction models at recovering missing drug-side effect associations in the held-out test set. Following a common practice in the literature,<sup>15,17,18</sup> we performed an ablation study. First, whenever possible, each method was trained using only the training matrix  $X$  without other side information (see first row in



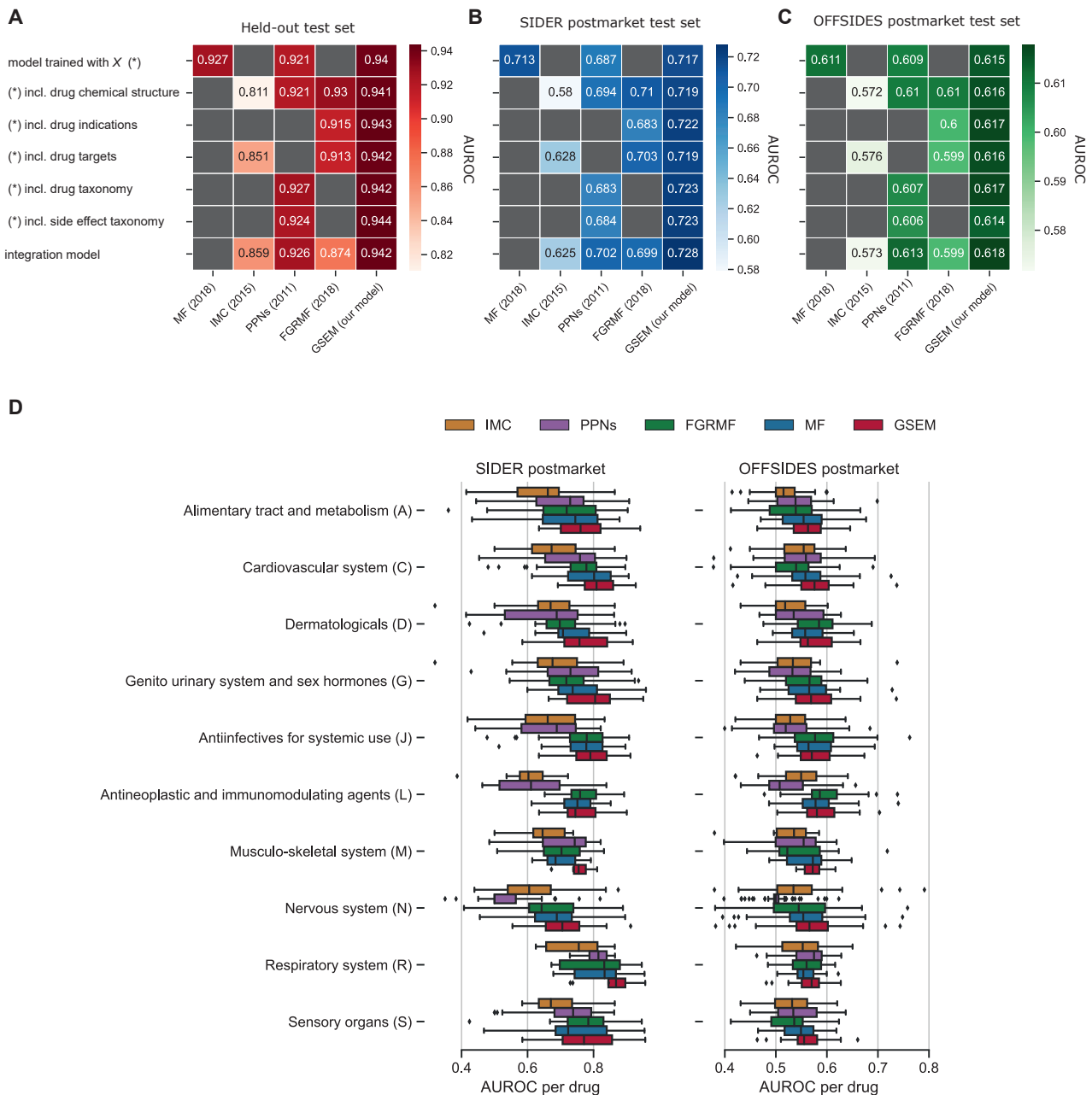
**Figure 2. Evaluation procedure**

(A) Drug side effect data were integrated from the SIDER 4.1 and OFFSIDES databases. They include a set of associations identified in clinical trials (red) and two sets of associations identified after the drugs entered the market: a postmarketing set from SIDER (blue) and OFFSIDES (green).

(B) The clinical trials association set was randomly split into training, validation, and test sets. Hyperparameters of each prediction model were tuned using the validation set. Each model was re-trained on the combined training and validation sets using optimal hyperparameters.

(C) Our test sets consisted of the held-out test set from the clinical trials set and the postmarketing sets from SIDER and OFFSIDES. Each positive set of associations was matched with a set of negatives twice their size, randomly selected.





**Figure 3. Performance evaluation on multiple drugs**

Each model (x axis) was trained with drug side effect associations obtained from clinical trials, without other information (first row, y axis), or in combination with one side information type at a time (chemical, indication, target, and taxonomy similarities): second to fifth rows. The methods that proposed a model to integrate multiple side information are indicated as the integration model in the last row of the heatmap. Area under the receiver operating curve (AUROC) is shown only for the side information types used in the original publications of each competitor. Gray cells represent N/A. The binary classification performance is shown for three independent test sets.

(A) (Red) Held-out test set containing other clinical trials side effects.

(B) (Blue) Postmarketing side effects from the SIDER database, containing side effects reported in package inserts that were identified after the drugs entered the market.

(C) (Green) Postmarketing side effects from the OFFSIDES database, containing statistically significant side effects from the Adverse Event Reporting System (AERS) surveillance database.

(D) Drug-specific performance according to its main category according to the Anatomical, Therapeutic, and Chemical (ATC) classification. (Left) AUROC in the SIDER postmarket test set; (right) AUROC in the OFFSIDES postmarket test set.

Figure 3A). Second, if possible, one side information at a time together with  $X$  was integrated into the model to assess its contribution to the overall performance (second to fifth rows in Figure 3A). In these experiments, we run each method with the side information types proposed in the original publications (see Methods S1). Finally, if the original publications proposed a way to integrate multiple information types (more than one) in their framework, we implemented them, and their performance is shown in the last row of Figure 3A. Notice that the GSEM, as proposed in Equations 3 and 2, is a model that allows for the integration of multiple types of heterogeneous information.

On the held-out test set with other side effects identified in clinical trials, the GSEM outperforms all the competitors by 1.4%–13.3%. Even when training GSEM using the training matrix  $X$  alone, i.e., without side information, the GSEM achieves 0.940 in terms of the AUROC. This baseline performance can be slightly improved using side information for drugs and side effects. Other methods, such as PPNs<sup>15</sup> and IMC,<sup>17</sup> also show a similar trend; therefore, side information should be used when available. In addition, we observed that while the competitors' performance is more sensitive to the specific choice of side information, the performance of the GSEM displays a small variability across information types. The mean and SD AUROCs in the held-out test set are  $0.9421 \pm 0.0012$  (GSEM) versus  $0.9079 \pm 0.0207$  (FGRMF),  $0.8405 \pm 0.0026$  (IMC), and  $0.9239 \pm 0.0212$  (PPNs). GSEM also consistently outperforms the competitors in terms of the AUPR (Figure S1).

We then tested our method in a more realistic scenario using a simulated prospective evaluation similar to the one used by Cami et al.<sup>15</sup> In this procedure, all side effects identified after the drugs entered the market were used as a test set (postmarket test sets in Figure 2B). Figures 3B and 3C show the prediction performance of the methods in postmarketing test sets. The GSEM outperforms the competitors by 1.5%–14.8% in the SIDER postmarket test set and by 0.7%–4.6% in the OFFSIDES postmarket test set.

Interestingly, the GSEM offers the best prediction performance in both prospective sets when combining all available side information. Following Cami et al.,<sup>15</sup> we further asked whether the performance of the models varies for drug- or side effect-specific categories. We performed a second evaluation where we used the best-performing models of each column of Figure 3A to analyze the performance of a specific group of drugs and side effects (see STAR Methods). Figure 3D shows the AUROC performance of the models for drug-specific anatomical categories according to their primary ATC classification. For most categories, the GSEM's mean AUROC was above 0.75 in the SIDER postmarket test set—we obtained the lowest AUROC performance for nervous system drugs (0.706) and the highest performance for respiratory system drugs (0.852). In the OFFSIDES test set, the mean AUROC was above 0.55 for all the categories. The performance of the models for the side effect-specific MedDRA category of disorders are shown in Figure S2.

### Distribution shifts in side effects reported before and after the drugs enter the market

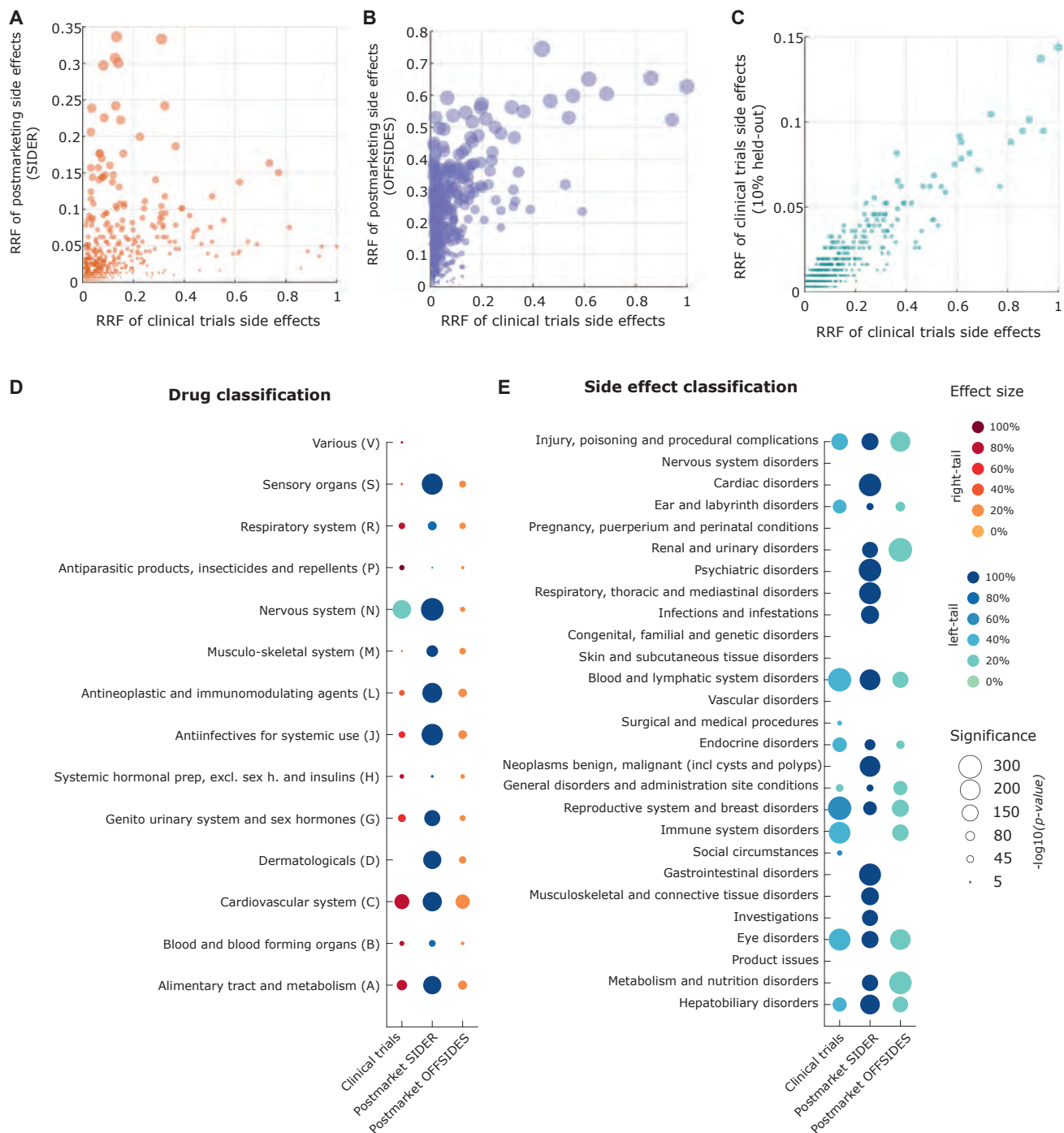
An important observation from Figures 3A–3C is that there is a considerable difference in AUROC performance when predicting

other side effects from clinical trials (GSEM AUROC of 0.944) versus postmarketing (GSEM AUROCs of 0.728 and 0.618 in the SIDER and OFFSIDES postmarket sets, respectively). These differences cannot be explained by the specific method used or the type of side information used in the integration. The differences in prediction performance prompted us to ask whether they can be explained by a distribution shift in side effect reports before and after the drug enters the market.

To analyze differences in reporting trends, we defined the ratio of reporting frequency (RRF) as the normalized count of drugs associated with a given side effect (see STAR Methods). The RRF reflects whether a side effect has been associated with many or few drugs in our dataset. For instance, nausea, a side effect reported on most drugs, has an RRF of 1.0, while eye infection, reported only on a few drugs, has an RRF of 0.011. We contrasted the RRF of each side effect computed using clinical trial associations versus postmarketing associations. Figures 4A and 4B show that side effects reported in clinical trials and postmarketing follow a different trend. A side effect reported on a small number of drugs in clinical trials (low RRF in the x axis) can be reported on many drugs in the postmarketing phase. This trend is even more prominent in the OFFSIDES postmarket set. For comparison, the expected trend without distribution shift is shown in Figure 4C for a held-out set from clinical trials associations (Pearson,  $\rho = 0.923$ ,  $p < 2.23 \times 10^{-308}$ ). Our results suggest differences in reporting trends between drug side effect associations reported in clinical trials and the postmarketing phase.

We further explored whether there are statistically significant differences in RRF values for drug anatomical classes and side effect disorder types. We grouped drugs by their main ATC classification and compared distributions of RRF values based on the known side effects reported in different sets (see STAR Methods). Figure 4D shows that for the majority of drug categories, the side effects that were reported in clinical trials tend to be biased toward frequently reported side effects except for nervous system drugs. Conversely, while the SIDER postmarket set tends to be more significant toward rarely reported side effects in clinical trials, the OFFSIDES set was more significant for frequently reported side effects. We repeated our statistical analysis by grouping side effects based on their main MedDRA category of disorders. Figure 4E shows that side effect categories are significant toward rarely reported side effects, i.e., low RRF values.

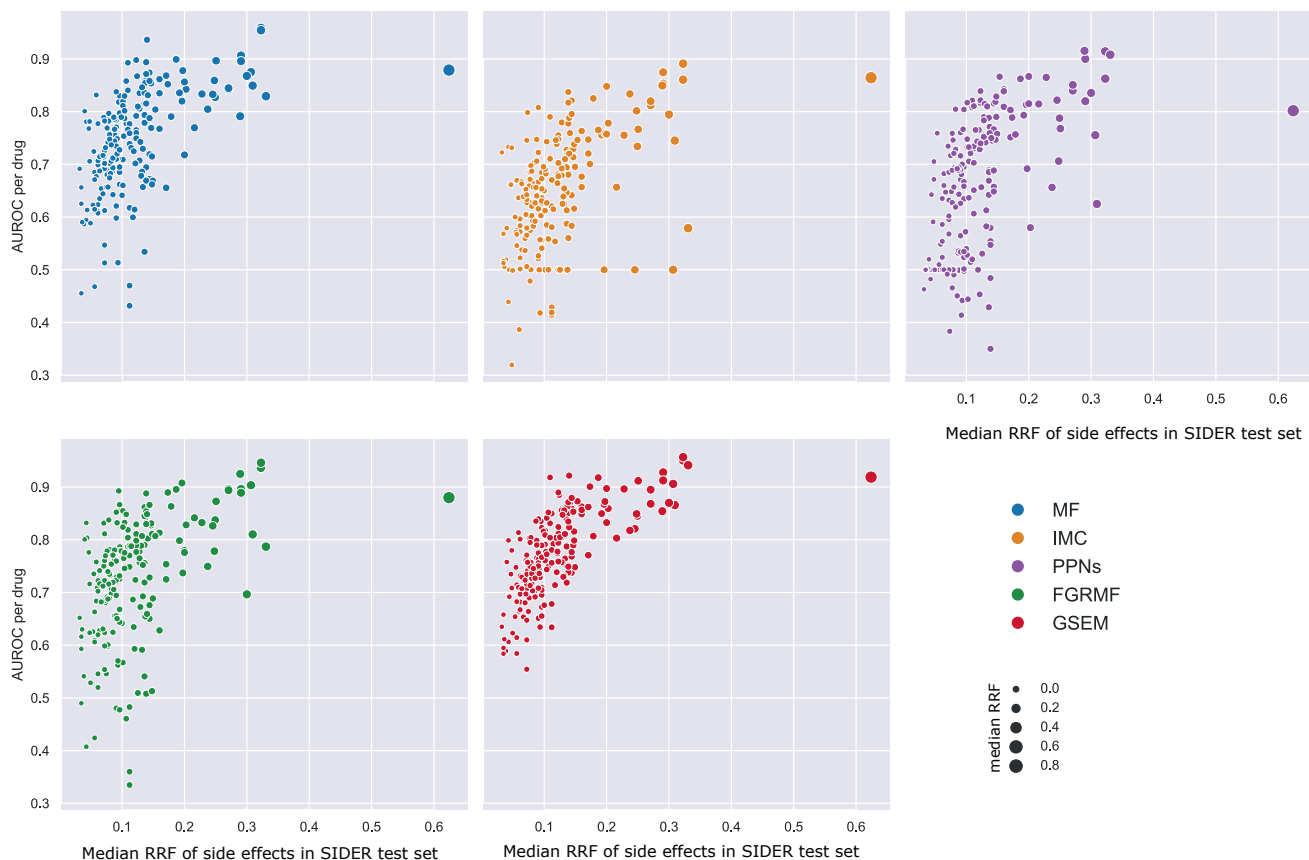
A fundamental assumption in machine learning is that the training and testing sets are drawn from the same underlying distribution.<sup>34</sup> Our analysis in Figure 4 shows that this is not the case for our problem. We hypothesized that the distribution shifts in side effect reports between clinical trials and postmarketing could explain the differences in prediction performance that we observed in Figures 3A–3C. It would imply a dependency between the AUROC performance and the RRF values of the side effects in the test set. To explore this dependency in more detail, we calculated AUROC values for single drugs on the SIDER postmarket test set. Figure 5 shows a correlation between prediction performance and the RRF values of the side effects we are trying to predict. A positive correlation is observed for all the methods, suggesting that each drug's



**Figure 4. Differences in the distribution of side effect reports in clinical trials and postmarketing drug development phases**

Side effect ratio of reporting frequency (RRF) is a normalized count of drugs associated with a given side effect. Each point represents a side effect, and the RRF values of side effects identified in clinical trials are compared against (A) the RRF of side effects identified in postmarketing as found in the SIDER database (Pearson,  $\rho = 0.377$ ,  $p < 5.1 \times 10^{-3.2}$ ); (B) the RRF of side effects identified in postmarketing as found in the OFFSIDES database (Pearson,  $\rho = 0.192$ ,  $p < 6.4 \times 10^{-9}$ ); and (C) the a held-out set (Pearson,  $\rho = 0.923$ ,  $p < 2.23 \times 10^{-308}$ ). The size of the circle is proportional to the RRF values.

(D and E) Statistical analysis of side effect RRF significance for (D) ATC group of drugs and (E) MedDRA-group of side effects. Only statistically significant associations are shown (one-tailed Wilcoxon rank-sum test with Benjamini-Hochberg adjusted significance,  $p < 0.05$ ). The circle size represents the significance ( $p$  value), and the color encodes the effect size of the association—the difference between the median in the group compared with the median of all drugs (or side effects). Colors separated the effect size to indicate whether the one-tailed significance was right-tailed (red palette) or left-tailed (blue palette).



**Figure 5. Dependency between prediction performance and side effect RRF value**

Each model generated scores by training with clinical trials' side effects and side information. Models were then assessed, for each drug, in their ability to identify the presence or absence of postmarketing side effects (SIDER postmarket test set) out of all the unknown side effects for the drug. Each dot in the figure represents an individual drug. The performance per drug is shown in the AUROC (y axis) versus the median RRF of the side effects in the test set (x axis). There is a direct correlation between the prediction performance of the each model and the median RRF value of the side effects in the test set: MF (Pearson correlation,  $\rho = 0.53$ ,  $p < 3.54 \times 10^{-16}$ ); IMC ( $\rho = 0.51$ ,  $p < 1.40 \times 10^{-14}$ ); PPNs ( $\rho = 0.55$ ,  $p < 2.85 \times 10^{-17}$ ); FGRMF ( $\rho = 0.45$ ,  $p < 2.50 \times 10^{-11}$ ); and GSEM ( $\rho = 0.68$ ,  $p < 4.11 \times 10^{-28}$ ). Each point represents a drug, and the circle's size is proportional to the median RRF.

prediction performance depends on the magnitude of the distribution shift.

Reported side effects in OFFSIDES have even lower RRF values than those in SIDER (see Figure S3), thus explaining the differences in AUROC performance between SIDER and OFFSIDES postmarket sets in Figures 3B and 3C, and Figure S4 shows that the AUROC per drug varies by category depending on the RRF values of the side effects in the postmarketing test sets.

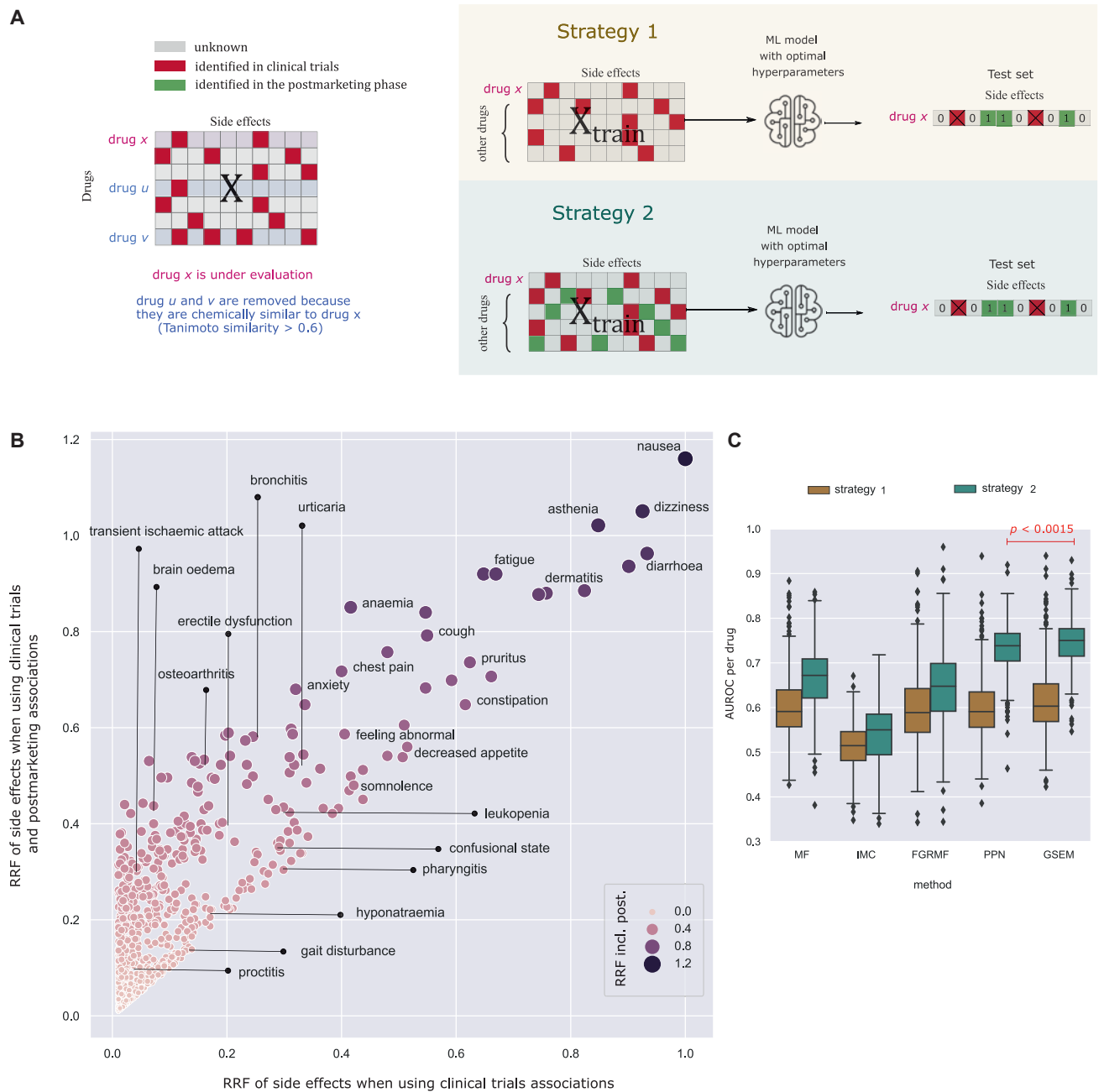
### A data integration technique to improve prediction performance

We propose a simple data integration technique to improve the prediction performance of side effect prediction models for individual drugs. Our idea is based on the observation that the effect of the distribution shift can be reduced if we integrate postmarketing data into the training matrix  $X$ . Figure 6B shows that the RRF values of specific side effects can be improved using postmarketing information in training.

Figure 6A illustrates our evaluation procedure for single drugs. For a given drug  $x$ , we used its clinical trials side effects for

training and its combined SIDER and OFFSIDES postmarketing side effects for testing. Then, we assessed the AUROC performance using two strategies that differ in the information used for the other drugs. The first strategy uses only side effect associations reported in clinical trials. The second strategy uses side effect associations reported in clinical trials and postmarketing. To prevent data leakage, we removed other chemically similar drugs from the training matrix  $X$  (see STAR Methods). Notice that for both strategies, we trained each method using the same set of optimal hyperparameters obtained in the validation set, as shown in Figure 2.

Figures 6B and 6C shows the AUROC performance of the side effect prediction models using strategies 1 and 2. The inclusion of the postmarketing side effects for the other drugs used for training dramatically affected the prediction performance for single drugs. The mean AUROC improved from 0.604 to 0.667 for MF; 0.512 to 0.537 for IMC; 0.596 to 0.650 for FGRMF; 0.60 to 0.733 for PNN; and 0.616 to 0.746 for the GSEM. Our method shows a 13% performance improvement using strategy 2.



**Figure 6. A data integration strategy for predicting postmarketing side effects for drugs in clinical trials**

(A) Evaluation procedure for single drugs to predict side effects identified after the drugs enter the market (postmarketing) using for training side effects identified in clinical trials. For a given drug  $x$ , we performed two evaluation strategies that change the set of associations used for the other drugs in  $X$ : (1) uses only clinical trials side effects and (2) uses clinical trials and postmarketing side effects. Side effects chemically similar to drug  $x$  were removed from the training matrix to avoid data leakage (illustrated as drugs  $u$  and  $v$ ).

(B) Comparison of side effect RRF values when using only clinical trials associations (x axis) and when also including also the postmarketing associations (y axis). Each point represents a side effect, and the circle's size is proportional to the RRF when including postmarketing side effects. Several side effect terms are indicated.

(C) Boxplots of the AUROC per drug on the combined postmarketing test sets using strategies 1 and 2. The distribution of AUROC values for the GSEM using strategy 2 is significantly better than that of the best competitor (PPN) (one-tailed Wilcoxon rank-sum test  $p < 0.0015$ ).

### Self-representations capture biological relationships

Two properties make the GSEM an interpretable and reproducible model. First, the GSEM is interpretable because the predicted score can be explained in terms of learned similarities between drugs and side effects. Second, the GSEM's solutions are reproducible because the learned solution is a globally optimal solution of its objective function. The GSEM overcomes the common problem of machine learning models that learn different solutions even when training the same model with a different random initialization, which is persistent in deep-learning models.<sup>35</sup>

The GSEM's predicted score for a drug  $i$  and side effect  $j$  can be written as follows:

$$\hat{X}_{ij} = \sum_{u \in \text{drugs known to cause side effect } j} H_{iu} + \sum_{v \in \text{side effects caused by drug } i} W_{vj}, \quad (\text{Equation 5})$$

where  $H$  and  $W$  are non-negative. The first term in Equation 5 contains the learned similarities between drug  $i$  and the drugs known to cause side effect  $j$ . The second term in Equation 5 contains the learned similarities between side effect  $j$  and the side effects known to be caused by drug  $i$ . If any of the individual terms in the sum is high, the prediction score  $\hat{X}_{ij}$  will be high because the model allows only for summation and not the subtraction of terms.

We hypothesized that the learned  $H$  can capture biological relationships between drugs. Following a similar procedure to Cheng et al.,<sup>36</sup> we assessed whether our drug similarity measure, defined as  $(H + H^T)/2$  (see STAR Methods), reflects known chemical, biological, and pharmacological relationships between drugs. To be sure that there is no information leakage, we trained the GSEM using all available clinical trials and postmarketing information (encoded in  $X$ ) but without any side information (i.e.,  $\mu_i = \alpha_j = 0 \forall i, j$ ) (see STAR Methods). We found that our drug similarity based on  $H$  correlates with chemical, indication, target, and ATC taxonomy similarities (Figure 7B). Interestingly, our drug similarity was also indicative that the drugs were pharmacologically similar (ATC taxonomy similarity above 0.05) or distinct (below 0.05). Our results suggest that the matrix  $H$  in our model could capture chemical, biological, and pharmacological relationships between drugs.

We also tested whether  $W$  could capture the anatomical/physiological relationships between side effect phenotypes, as defined by the MedDRA taxonomy similarity (see STAR Methods). We defined side effect similarities based on  $W$  as  $(W + W^T)/2$  (see STAR Methods). We found that the side effect similarities based on  $W$  correlate with the MedDRA taxonomy similarity (Figure 7B, bottom). We observed that phenotypically similar side effects tend to have similar self-representations. The similarity also indicates whether side effects are anatomically/physiologically similar (MedDRA taxonomy similarity above 0.05) or distinct (below 0.05).

To showcase how the learned matrices allow for interpretability, we explored the weights in  $W$  for two side effects: (1) myocardial infarction (MI), which has been associated with the withdrawal of many drugs from the market,<sup>4</sup> and (2) blurred vision. Figure 7A shows a diagram of the side effects that are more similar to MI and blurred vision based on the weights in  $W$ . We observed that MI is very similar to other vascular-related disorders, including

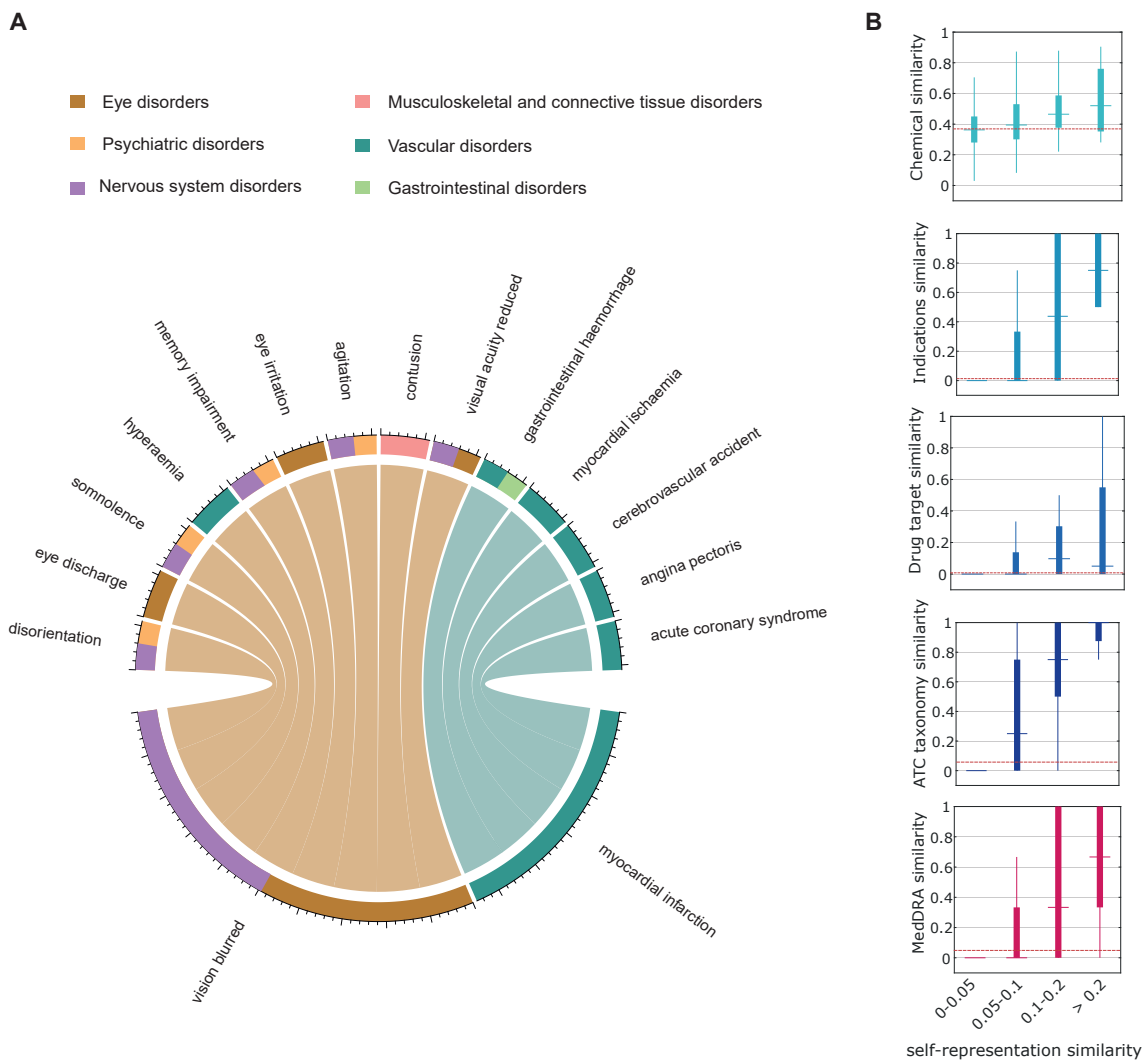
angina pectoris, which has been shown to appear prior to MI.<sup>37</sup> 46 drugs in our dataset are known to be associated with both angina pectoris and MI, which might explain the learned association. On the other hand, blurred vision, which is classified in MedDRA as both an eyes and nervous system disorder, is also very similar to other related conditions, including psychiatric disorders. The learned matrix  $W$  allows for a transparent inspection of how the model arrived at a given prediction. If a drug is known to induce MI, our model predicts that the drug might also induce similar side effects, as shown in Figure 7A.

### DISCUSSION

Here, we introduced the GSEM, a computational approach for predicting the side effects of drugs during clinical drug development. Instead of waiting for postmarketing observational evidence to be accumulated, our framework can be used to assist drug safety professionals in the identification of side effects during drug clinical trials. To show this, we trained the models with side effects identified in clinical trials and tested them to predict side effects identified in the postmarketing phase. To our knowledge, this is the first attempt to predict the presence or absence of side effects for drugs with a small number of side effects identified in clinical trials. Our framework can be used together with our recent approach to predict the frequencies of drug side effects in patients.<sup>28</sup> These tools can be helpful in the early detection of rare side effects that cannot be effectively captured in small-size clinical trials.

Our analysis indicated that predicting side effects that were identified after the drugs entered the market is difficult when training only with side effects identified during clinical trials. Part of this difficulty lies in the differences in the distribution of side effects reported in clinical trials and in postmarketing. Scarcely reported side effects during clinical trials tend to be highly reported in postmarketing, thus explaining the models' difficulty at predicting them. We further studied this issue by analyzing the dependency between the number of drugs associated with a side effect (RRF value) and the prediction performance of machine learning models (see Figure 5). Our experiments showed that the prediction performance of the models heavily depended on the RRF value of the side effects we were aiming to predict. Strikingly, improving the RRF value of each side effect by adding information from postmarketing reports was more critical for improving the prediction of postmarketing side effects than the use of any drug or side effect features.

The problem of distribution shift in side effect reports is deeply connected to the intrinsic distributional properties of drug side effects. In a previous study,<sup>28</sup> we have shown that drug side effect reports follow a long-tailed distribution. The distribution can be summarized in a Pareto 80/30 rule, where 80% of the associations come from 30% of the side effects.<sup>28</sup> Unfortunately, this means that the amount of labeled information (captured by RRF), vital for machine learning models, varies per side effect, following an almost exponential distribution. It would be essential to consider the dependency between prediction performance and side effect RRF when evaluating computational models that aim to predict drug side effects.



**Figure 7. Self-representations capture chemical, biological, and pharmacological relationships**

(A) Diagram representing how vision blurred and MI (bottom) are self-represented with other side effects (top). Only side effects with self-representations weights above 0.05 are shown. The thickness of the connections is proportional to the self-representation weights in  $W$ . The colors in the outer circle represent the disorder category of the side effect according to the Medical Dictionary for Regulatory Activities (MedDRA) terminology.

(B) The interplay between the drug self-representation similarity and four types of drug-drug similarities: chemical, indications, target, and ATC taxonomy. The bottom figure shows the interplay between the side effect self-representation similarity and the MedDRA taxonomy similarity. Mean values of chemical (mean similarity of 0.3689), indications (0.0134), drug target (0.0076), ATC taxonomy (0.0576), and MedDRA taxonomy (0.0488) similarities are shown as dashed horizontal lines.

An innovative aspect of our algorithm is that it learns similarities between drugs (matrix  $H$ ) and between side effects (matrix  $W$ ). Our model is fundamentally different from previous side effect prediction models. A PPN<sup>15</sup> is a network-based method that builds topological features from the bipartite drug-side effect graph. The graph is obtained when connecting the nodes representing drugs to the set of nodes representing side effects. PPNs also integrate chemical, taxonomic, and biological features and then use a logistic regression model to predict. MF<sup>16</sup> is a matrix decomposition model that learns a low-dimensional feature vector for each drug and a low-dimensional feature vector for each side effect such that the dot product between the vectors models

an entry in  $X$ . It amounts to a low-rank approximation of  $X$ . Similarly, FGRMF<sup>18</sup> uses several low-rank approximation models for each drug side information graph that are integrated into the model using the smoothness constraint.<sup>24–26</sup> The final FGRMF score is the probability given the logistic regression that combines the scores of the individual low-rank models. Finally, IMC<sup>17</sup> is an IMC model that integrates drugs and side effect features in the matrix decomposition model. A detailed description of the mathematical formulation of each competitor, together with their implementation and optimization, can be found in Methods S1.

GSEM builds upon the recent development of high-rank matrix completion based on self-expressive models (SEM)<sup>38</sup> and sparse

linear methods,<sup>39</sup> as well as the recent trend of deep learning on graphs.<sup>26,40,41</sup> SEMs represent data points, e.g., drugs, approximately as a linear combination of a few other data points. Elhamifar<sup>38</sup> proposed SEMs as a framework for simultaneously clustering and completing high-dimensional data lying in the union of low-dimensional subspaces. It has been shown that SEMs generalize over standard low-rank matrix completion models,<sup>42,43</sup> which might explain why the GSEM outperforms previous approaches that have been proposed to predict drug side effects based on low-rank matrix decomposition.<sup>16–18</sup> Self-representations naturally allow the integration of graph-based information about drugs or side effects. Our model is also related to non-negative MF (NMF).<sup>27,44</sup> They differ, however, in two main aspects. First, while NMF learns two low-rank matrices to represent the input data, the GSEM learns a single null-diagonal matrix that allows for a high-rank matrix.<sup>38</sup> Second, while the NMF objective function is non-convex, we proved that our objective function is convex and that our algorithm converges to a globally optimal solution.

Our framework could be easily applied to proprietary datasets of drug side effects by following our procedure illustrated in Figure 2. The GSEM is fast to run, and its prediction performance is robust to the specific choice of hyperparameters (see our analysis in Figure S5). Applying our model for a compound undergoing clinical trials is as easy as adding the new compound information in a new row in *X*. We started investigating the potential of the GSEM for drug repositioning,<sup>45</sup> and we envision applying our algorithm to other open problems in biology, chemistry, and medicine, such as drug target prediction<sup>46</sup> or antiviral drug effect prediction.<sup>47</sup> To assist scientists working in clinical drug development in their difficult task, we provide the code to run our algorithm (<https://github.com/paccanarolab/GSEM>), the predictions for the 505 drugs used in our study (supplementary dataset 4 in Galeano and Paccanaro<sup>48</sup>), and the learned matrices that can help to interpret the predictions (supplementary datasets 5 and 6 in Galeano and Paccanaro<sup>48</sup>).

Whenever machine learning models support high-stakes decisions, it is desirable to have inherently interpretable models.<sup>49</sup> We have shown that the learned matrices in our model capture biological and pharmacological relationships between drugs and physiological relationships between side effect phenotypes. However, the medical, biological, or pharmacological interpretation of the relationships requires expert biological and medical knowledge. In the supplemental information, we also discussed the differences between the interpretability capabilities of the GSEM and our latent factor model for predicting the frequencies of drug side effects<sup>28</sup> (see Methods S3).

### Limitations of the study

We run our method only for drugs with at least five side effects identified in clinical trials. A limitation of expanding our analysis is the lack of standardized datasets that classify side effects depending on the phase of the clinical trial in which it was identified.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets
  - Side effect ratio of reporting frequency (RRF)
  - Similarities in side information graphs
  - Model selection and evaluation for multiple drugs
  - Performance evaluation for single drugs
  - Multiplicative learning algorithm
- QUANTIFICATION AND STATISTICAL ANALYSIS

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.crmeth.2022.100358>.

### ACKNOWLEDGMENTS

We thank Mateo Torres, Suzana de Siqueira Santos, Ruben Jimenez, Santiago Noto, and Philip Ovington for useful discussions. D.G. was supported by the 2022 KBR SMS SDI Lymphoma grant and the US Air Force grant contract no. FA8075-16-D-0010, task order FA8075-18-F-1690 Explainable Artificial Intelligent Applications within Integrated Dynamic Visualization Environment, and the Facultad de Ingenieria - UNA. A.P. was supported by Biotechnology and Biological Sciences Research Council (<https://bbsrc.ukri.org/>) grant numbers BB/K004131/1, BB/F00964X/1, and BB/M025047/1; Medical Research Council (<https://mrc.ukri.org>) grant number MR/T001070/1; Consejo Nacional de Ciencia y Tecnología Paraguay (<https://www.conacyt.gov.py/>) grants numbers 14-INV-088, PINV15-315, and PINV20-337; National Science Foundation Advances in Bio Informatics (<https://www.nsf.gov/>) grant number 1660648; Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro grant number E-26/201.079/2021 (260380); and Fundação Getulio Vargas.

### AUTHOR CONTRIBUTIONS

Conceptualization, D.G.; methodology, D.G. and A.P.; investigation, D.G.; formal analysis, D.G.; writing – original draft, D.G.; writing – review & editing, D.G. and A.P.; software, D.G.; supervision, A.P.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 12, 2022

Revised: September 8, 2022

Accepted: November 11, 2022

Published: December 7, 2022

### REFERENCES

1. GBD 2016 Causes of Death Collaborators; Abajobir, A.A., Abbafati, C., Abbas, K.M., Abd-Allah, F., Abera, S.F., Aboyans, V., Adetokunboh, O., Afshin, A., Agrawal, A., et al. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the global burden of disease study 2016. *Lancet* 390, 1151–1210.
2. Sunshine, J.E., Meo, N., Kassebaum, N.J., Collison, M.L., Mokdad, A.H., and Naghavi, M. (2019). Association of adverse effects of medical treatment with mortality in the United States: a secondary analysis of the global burden of diseases, injuries, and risk factors study. *JAMA Netw. Open* 2, e187041.



3. Martin, L., Hutchens, M., Hawkins, C., and Radnov, A. (2017). How much do clinical trials cost? *Nat. Rev. Drug Discov.* *16*, 381–382.
4. Onakpoya, I.J., Heneghan, C.J., and Aronson, J.K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med.* *14*, 10.
5. Ho, T.-B., Le, L., Thai, D.T., and Taewijit, S. (2016). Data-driven approach to detect and predict adverse drug reactions. *Curr. Pharm. Des.* *22*, 3498–3526.
6. Bolland, M.R., Jacunski, A., Lorberbaum, T., Romano, J.D., Moskovitch, R., and Tatonetti, N.P. (2016). Systems biology approaches for identifying adverse drug reactions and elucidating their underlying biological mechanisms. *Wiley Interdiscip. Rev. Syst. Biol. Med.* *8*, 104–122.
7. Yamanishi, Y., Pauwels, E., and Kotera, M. (2012). Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model.* *52*, 3284–3292.
8. Filiri, A.F., Loging, W.T., Thadeio, P.F., and Volkmann, R.A. (2005). Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* *1*, 389–397.
9. Atias, N., and Sharan, R. (2011). An algorithmic framework for predicting side effects of drugs. *J. Comput. Biol.* *18*, 207–218.
10. Lounkine, E., Keiser, M.J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J.L., Lavan, P., Weber, E., Doak, A.K., Côté, S., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* *486*, 361–367.
11. Poleksic, A., and Xie, L. (2018). Predicting serious rare adverse reactions of novel chemicals. *Bioinformatics* *34*, 2835–2842.
12. LaBute, M.X., Zhang, X., Lenderman, J., Bennion, B.J., Wong, S.E., and Lightstone, F.C. (2014). Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PLoS One* *9*, e106298.
13. Scheiber, J., Chen, B., Milik, M., Sukuru, S.C.K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., et al. (2009). Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* *49*, 308–317.
14. Zhou, H., Gao, M., and Skolnick, J. (2015). Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Sci. Rep.* *5*, 11090.
15. Cami, A., Arnold, A., Manzi, S., and Reis, B. (2011). Predicting adverse drug events using pharmacological network models. *Sci. Transl. Med.* *3*, 114ra127.
16. Galeano, D., and Paccanaro, A. (2018). A recommender system approach for predicting drug side effects. In 2018 International Joint Conference on Neural Networks (IJCNN) (IEEE), pp. 1–8.
17. Li, R., Dong, Y., Kuang, Q., Wu, Y., Li, Y., Zhu, M., and Li, M. (2015). Inductive matrix completion for predicting adverse drug reactions (adrs) integrating drug–target interactions. *Chemometr. Intell. Lab. Syst.* *144*, 71–79.
18. Zhang, W., Liu, X., Chen, Y., Wu, W., Wang, W., and Li, X. (2018). Feature-derived graph regularized matrix factorization for predicting drug side effects. *Neurocomputing* *287*, 154–162.
19. Bean, D.M., Wu, H., Iqbal, E., Dzahini, O., Ibrahim, Z.M., Broadbent, M., Stewart, R., and Dobson, R.J.B. (2018). Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.* *8*, 4284.
20. Knepper, T.C., and McLeod, H.L. (2018). When will clinical trials finally reflect diversity? *Nature* *557*, 157–159.
21. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The sider database of drugs and side effects. *Nucleic Acids Res.* *44*, D1075–D1079.
22. Ng, A.Y. (2004). Feature selection,  $l_1$  vs.  $l_2$  regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine Learning (ACM), p. 78.
23. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Royal Statistical Soc. B* *67*, 301–320.
24. Ma, H., Zhou, D., Liu, C., Lyu, M.R., and King, I. (2011). Recommender systems with social regularization. In Proceedings of the fourth ACM international conference on Web search and data mining (ACM), pp. 287–296.
25. Kalofolias, V., Bresson, X., Bronstein, M., and Vandergheynst, P. (2014). Matrix completion on graphs. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1408.1717>.
26. Monti, F., Bronstein, M., and Bresson, X. (2017). Geometric matrix completion with recurrent multi-graph neural networks. In Advances in Neural Information Processing Systems, pp. 3697–3707.
27. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* *401*, 788–791.
28. Galeano, D., Li, S., Gerstein, M., and Paccanaro, A. (2020). Predicting the frequencies of drug side effects. *Nat. Commun.* *11*, 4575.
29. Tatonetti, N.P., Ye, P.P., Daneshjou, R., and Altman, R.B. (2012). Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* *4*, 125ra31.
30. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* *46*, D1074–D1082.
31. Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., et al. (2017). The drug repurposing hub: a next-generation drug library and information resource. *Nat. Med.* *23*, 405–408.
32. MDL Information Systems/Symyx (1984). MACCS-II (MDL Information Systems/Symyx).
33. Landrum, G. (2013). Rdkit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling (Sourceforge).
34. Bishop, C.M. (2006). *Pattern Recognition and Machine Learning* (Springer).
35. Hinton, G. (2018). Deep learning—a technology with the potential to transform health care. *JAMA* *320*, 1101–1102.
36. Cheng, F., Kovács, I.A., and Barabási, A.L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* *10*, 1197.
37. Behar, S., Reicher-Reiss, H., Abinader, E., Agmon, J., Friedman, Y., Barzilai, J., Kaplinsky, E., Kauli, N., Kishon, Y., Palant, A., et al. (1992). The prognostic significance of angina pectoris preceding the occurrence of a first acute myocardial infarction in 4166 consecutive hospitalized patients. *Am. Heart J.* *123*, 1481–1486.
38. Elhamifar, E. (2016). High-rank matrix completion and clustering under self-expressive models. In Advances in Neural Information Processing Systems, pp. 73–81.
39. Ning, X., and Karypis, G. (2011). Slim: sparse linear methods for top-n recommender systems. In Data Mining (ICDM), 2011 IEEE 11th International Conference (IEEE), pp. 497–506.
40. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* *34*, 18–42.
41. Hamilton, W.L., Ying, R., and Leskovec, J. (2017). Representation learning on graphs: methods and applications. *IEEE Data Eng. Bull.* *40*, 52–74.
42. Fan, J., and Chow, T.W. (2017). Matrix completion by least-square, low-rank, and sparse self-representations. *Pattern Recogn.* *71*, 290–305.
43. Wang, Y., and Elhamifar, E. (2018). High rank matrix completion with side information. In Thirty-Second AAAI Conference on Artificial Intelligence.
44. Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pp. 556–562.
45. Frasca, F., Galeano, D., Gonzalez, G., Laponogov, I., Veselkov, K., Paccanaro, A., and Bronstein, M.M. (2019). Learning interpretable disease

- self-representations for drug repositioning. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1909.06609>.
46. Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J.K., Chupakhin, V., Ceulemans, H., and Moreau, Y. (2017). Macau: scalable bayesian factorization with high-dimensional side information using mcmc. In 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP) (IEEE), pp. 1–6.
  47. Santos, S.d.S., Torres, M., Galeano, D., Sánchez, M.D.M., Cernuzzi, L., and Paccanaro, A. (2022). Machine learning and network medicine approaches for drug repositioning for covid-19. *Patterns* 3, 100396.
  48. Galeano, D., and Paccanaro, A.. (2022). Machine Learning Prediction of Side effects for Drugs in Clinical Trials - Galeano and Paccanaro. *Mendley*. <https://doi.org/10.17632/3z7c4r52n3.1>.
  49. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215.
  50. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J., and Bork, P. (2008). Drug target identification using side-effect similarity. *Science* 321, 263–266.

## Article

# TriNet: A tri-fusion neural network for the prediction of anticancer and antimicrobial peptides

Wanyun Zhou,<sup>1,5</sup> Yufei Liu,<sup>1,5</sup> Yingxin Li,<sup>2</sup> Siqi Kong,<sup>1</sup> Weilin Wang,<sup>1</sup> Boyun Ding,<sup>1</sup> Jiyun Han,<sup>3</sup> Chaozhou Mou,<sup>3</sup> Xin Gao,<sup>4,\*</sup> and Juntao Liu<sup>3,6,\*</sup>

<sup>1</sup>SDU-ANU Joint Science College, Shandong University (Weihai), Weihai 264209, China

<sup>2</sup>School of Mechanical, Electrical & Information Engineering, Shandong University (Weihai), Weihai 264209, China

<sup>3</sup>School of Mathematics and Statistics, Shandong University (Weihai), Weihai 264209, China

<sup>4</sup>Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

<sup>5</sup>These authors contributed equally

<sup>6</sup>Lead contact

\*Correspondence: xin.gao@kaust.edu.sa (X.G.), juntaosdu@126.com (J.L.)

<https://doi.org/10.1016/j.patter.2023.100702>

**THE BIGGER PICTURE** In drug discovery, the importance of antimicrobial peptides is increasing as multi-drug-resistant microbes continue to emerge. In addition, there is a growing clinical interest in anticancer peptides for the treatment of drug-resistant cancer cells. The cost of traditional wet lab experiments to identify such peptides can be significantly reduced by using computational methods that utilize artificial intelligence. In this study, we developed a deep-learning framework called TriNet for the accurate and rapid identification of anticancer and antimicrobial peptides. Benchmarking studies demonstrate that TriNet performs with extensive adaptability and effectiveness in identifying anticancer and antimicrobial peptides. In this work, TriNet is improved through the appropriate constructions of peptide features, the tri-fusion neural network, and the TVI training method. Further refinement may lead to an effective tool for guiding cancer treatment and antibiotic drug design.



**Development/Pre-production:** Data science output has been rolled out/validated across multiple domains/problems

## SUMMARY

The accurate identification of anticancer peptides (ACPs) and antimicrobial peptides (AMPs) remains a computational challenge. We propose a tri-fusion neural network termed TriNet for the accurate prediction of both ACPs and AMPs. The framework first defines three kinds of features to capture the peptide information contained in serial fingerprints, sequence evolutions, and physicochemical properties, which are then fed into three parallel modules: a convolutional neural network module enhanced by channel attention, a bidirectional long short-term memory module, and an encoder module for training and final classification. To achieve a better training effect, TriNet is trained via a training approach using iterative interactions between the samples in the training and validation datasets. TriNet is tested on multiple challenging ACP and AMP datasets and exhibits significant improvements over various state-of-the-art methods. The web server and source code of TriNet are respectively available at <http://liulab.top/TriNet/server> and <https://github.com/wanyunzh/TriNet>.

## INTRODUCTION

The dramatic increase in antimicrobial resistance poses a severe threat to public health globally.<sup>1</sup> Due to the misuse or overuse of antibiotic drugs, some bacterial pathogens generate resistance to antimicrobials, which has adverse ef-

fects on disease treatments.<sup>2,3</sup> Consequently, the discovery of alternative therapies for combating infections caused by multidrug-resistant bacteria is urgently needed.<sup>4</sup> One promising strategy is to perform therapy based on antimicrobial peptides (AMPs), which can help reduce the likelihood of resistance emergence.<sup>5</sup>



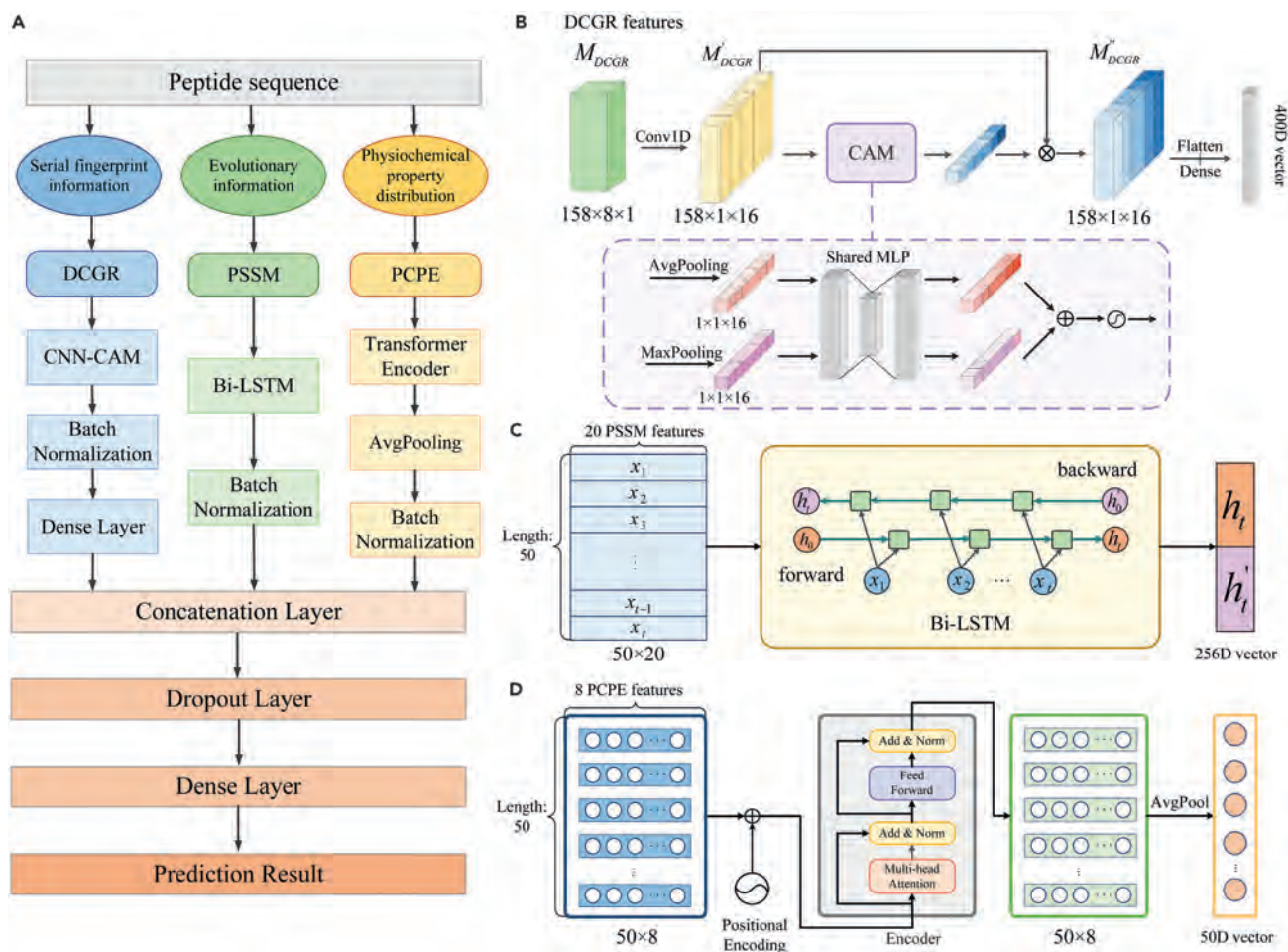
A great number of known AMPs are small molecules with negligible toxicity and broad spectra of activity against bacteria, fungi, viruses, and even cancer cells.<sup>6,7</sup> Anticancer peptides (ACPs) are a specific class of AMPs that can control cancer cell resistance to anticancer drugs.<sup>8</sup> Similar to most AMPs, ACPs with cations can engage electrostatically with the anionic membranes of cancer cells and kill them without destroying normal cells.<sup>9</sup> In recent years, AMPs, including ACPs, have been widely used for clinical applications in a variety of disease therapies.<sup>10–13</sup> Accordingly, the effective identification of peptides with biological activity is crucial for developing candidate drugs. Various experimental and computational methods have been developed. Traditional wet experiments are often expensive and time consuming; hence, the development of reliable computational methods is urgently needed. With the development of artificial intelligence, an increasing number of computational methods based on machine learning have been proposed. For those methods, the extraction of effective peptide sequence features is the critical first step. In recent decades, researchers have explored various algorithms for extracting features from the compositional and distribution information of amino acid sites, and other approaches take advantage of the physicochemical properties that set AMPs or ACPs apart from other peptide sequences. In addition, binary profile features (BPFs),<sup>14</sup> amino acid composition (AAC), and dipeptide composition (DPC)<sup>15</sup> are also widely employed. Based on AAC, Chou<sup>16</sup> proposed the PseAAC model to preserve sequence order information. Wei et al.<sup>17</sup> proposed an adaptive skip DPC (ASDC) method for enriching DPC features. The compositional-transition-distribution (CTD) algorithm proposed by Dubchak et al.<sup>18</sup> clusters 20 amino acids into three groups based on specific physicochemical properties and summarizes 21 descriptors containing composition, transition, and distribution information, which can better describe the global compositions of the physicochemical properties of amino acids in peptide sequences.

With the tremendous development of deep learning, in addition to the use of traditional machine learning algorithms, such as support vector machines (SVMs), random forests (RFs), and extreme gradient boosting (XGBoost),<sup>19</sup> deep-learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are increasingly being employed by researchers to identify functional peptides. For the prediction of AMPs, Veltri et al.<sup>20</sup> transformed amino acid residues into 128-dimensional vectors via an embedding layer and combined a convolution layer and a recurrent layer to capture potential sequence information. Su et al.<sup>21</sup> used multiscale convolutional layers with different filter lengths to capture multiscale motifs in peptide sequences. Fu et al.<sup>22</sup> proposed a deep neural network (DNN) model called ACEP using convolutional layers and an attention mechanism to fuse the feature tensors generated by a learnable sequence encoding model. For ACP detection, the classical model is the long short-term memory (LSTM) neural network-based deep-learning framework developed by Yi et al. called ACP-DL.<sup>23</sup> Ahmed et al.<sup>24</sup> proposed a multi-headed deep CNN model, ACP-MHCNN, for extracting features from different sources of information, such as physicochemical properties and evolutionary information, using parallel CNNs for ACP prediction. Wang et al.<sup>25</sup> proposed a hybrid CNN-LSTM model termed CL-ACP that applies a CNN to focus on

local information and an LSTM to extract the dependencies of residues. Lv et al.<sup>26</sup> used two kinds of sequence-embedding technologies, SSA and UniRep, related to DNNs based on LSTM to complete classification tasks.

The existing methods for predicting ACPs and AMPs mainly have the following shortcomings. In terms of feature extraction, the features of peptide sequence residues are usually extracted in a one-by-one manner in most existing predictors. Thus, the global information on peptide sequences cannot be captured. Methods such as ACEP, which uses attention scores to capture relationships across peptides, should be adopted. In addition, in existing methods, several physicochemical properties are usually selected directly from hundreds of properties,<sup>14</sup> which may result in serious redundancy or low quality of the chosen properties. In terms of the design of neural networks for processing extracted features, many models fail to design specific neural networks based on the properties of different features and even apply the same or similar neural network architectures to process different kinds of features. Without effective peptide feature processing, the performance of existing methods still has plenty of room for improvement. In terms of neural network training, the training and validation sets are randomly separated in traditional training approaches. Thus, there is no guarantee that hard samples (samples that are very likely to be wrongly predicted) are well trained, since they may be totally split into the validation set with no or only a few samples in the training set. In recent years, several partition approaches, including SPXY,<sup>27</sup> Rank-KS,<sup>28</sup> and SPXYE,<sup>29</sup> have been proposed to split training and validation sets. The core idea of these methods is to repeatedly select samples with the maximal distance until a predefined number of samples is obtained. Then, the selected and remaining samples are regarded as training and validation sets, respectively. However, in all of these methods, the separations are performed prior to training, ignoring the possibility that different neural networks have different hard samples. Therefore, more appropriate feature extraction methods, neural networks, and separations of training and validation sets are urgently needed to improve the identification of ACPs and AMPs.

In this study, we introduce TriNet, a tri-fusion neural network for ACP and AMP prediction (see Figure 1 for the workflow of TriNet). (1) TriNet is designed based on the assumption that whether a peptide is an ACP or AMP should be determined by multiple pieces of information and their effective fusion. (2) In addition to the frequently used position-specific scoring matrix (PSSM) feature, TriNet introduces another two features for representing the information contained in the serial fingerprint and physicochemical properties of a peptide sequence. (3) TriNet employs three parallel networks, a channel attention module (CAM) based on convolutional layers (for processing serial fingerprint features), a bidirectional LSTM network (Bi-LSTM; for processing the sequence evolution features), and an encoder module (for processing physicochemical property features), attempting to effectively fuse the above three kinds of features. (4) Different from traditional neural network training methods, TriNet is trained by a training approach termed TVI to achieve a better training effect, which is achieved by iterative interactions between the samples in the training and validation datasets to generate more appropriate training and validation sets based on the biases of neural networks.



**Figure 1. Overall structure of TriNet**

(A) Flowchart of the proposed TriNet model.

(B) Architecture of the DCGR-CNN-CAM mechanism. First, a matrix  $M_{DCGR}$  containing serial fingerprint information is fed into a convolutional layer, and a feature map  $M'_{DCGR}$  is generated. Then, a CAM layer is conducted on  $M'_{DCGR}$  to obtain the channel weights, and the weight-assigned feature map  $M''_{DCGR}$  is flattened and passed through a dense layer.

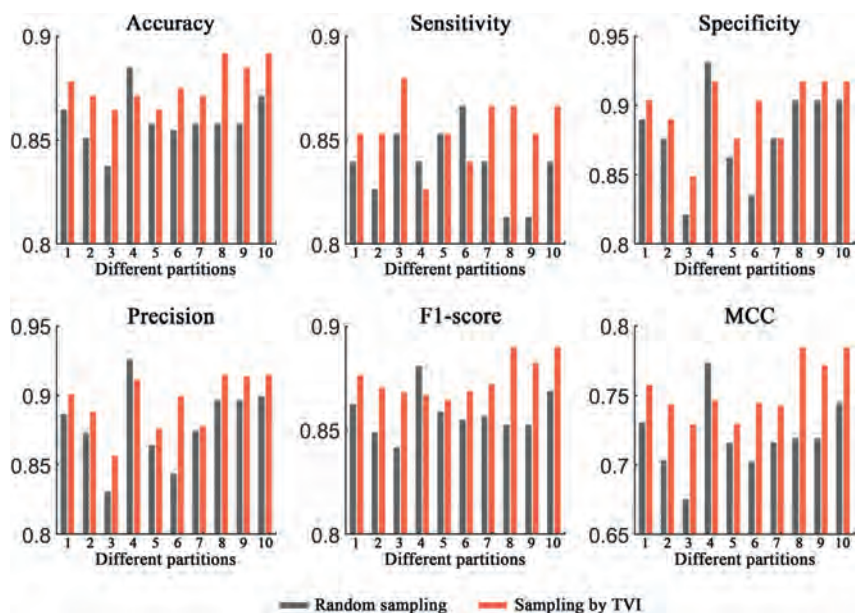
(C) Architecture of the PSSM-Bi-LSTM module. Given a feature matrix  $M_{PSSM}$ , a Bi-LSTM network is applied to process the sequence evolution features.

(D) Architecture of the PCPE-encoder module. The feature matrix  $M_{PCPE}$  is first fed into the encoder block of a transformer by using positional encoding, and then average pooling is applied after the encoder module.

We benchmarked TriNet on multiple challenging ACP and AMP datasets by using both cross-validation and independent testing, and the results showed that the proposed framework achieved substantially improved performance over that of other ACP/AMP prediction tools. In addition, we fully evaluated the effectiveness of the TVI training method for the prediction of both ACPs and AMPs, and in multiple other network models, and found that TVI effectively reconstructed the most appropriate training and validation sets based on the biases of a given neural network. Finally, we tested the effectiveness of the three proposed features and network structures on all six datasets, and the results clearly demonstrated the extensive adaptability and effectiveness of the extracted features and the network structures. TriNet has been proven to be very sensitive in detecting ACPs and AMPs, demonstrating its great potential for guiding the development of small-peptide drugs targeting cancer cells or other pathogens, such as bacteria, fungi, and viruses.

## RESULTS

TriNet is a framework for predicting ACPs/AMPs based on peptide sequences by effectively fusing the information contained in the serial fingerprints, sequence evolutions, and physicochemical properties of peptide sequences and then training the network with a training method called TVI. We first tested the effectiveness of the TVI training method by comparing it with the traditional training method (random sampling). Then, we evaluated the performance of TriNet on a diverse set of challenging datasets and compared it with six other ACP prediction algorithms, ACP-DL,<sup>23</sup> MHCNN,<sup>24</sup> iACP-DRLF,<sup>26</sup> CL-ACP,<sup>25</sup> DeepACPpred,<sup>30</sup> and AntiCP 2.0,<sup>31</sup> as well as six AMP prediction algorithms, DNN,<sup>20</sup> APIN,<sup>21</sup> ACEP,<sup>22</sup> CAMP-RF, CAMP-SVM, and CAMP-ANN.<sup>32</sup> Finally, we analyzed the effectiveness of the extracted features as well as the structures of TriNet. In this study, the accuracy, sensitivity, specificity, precision, F1 score, and Matthews



**Figure 2. Performance comparison between the traditional training approach and the TVI method on the ACP740 dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

correlation coefficient (MCC) metrics were employed as evaluation criteria (see the experimental procedures).

### Performance evaluation of the TVI training method

In this section, two ACP datasets (ACP740 and ACPmain) and an AMP dataset (Xiao dataset) were used to evaluate the effectiveness of the TVI training method, and the process was as follows. For ACP740, 20% of the ACPs and non-ACPs were randomly selected as the fixed test set, and the remaining 80% of the samples were then randomly separated into a training set (containing 473 samples) and a validation set (containing 119 samples). The random separation of the training and validation sets was performed 10 times, and the 10 different pairs of training and validation sets produced were used for network training and validation, respectively. Then, different trained models were evaluated on the test set, and the results showed that the network models demonstrated obvious biases on different separations of the training and validation sets (see Figures 2, 3, and 4; random sampling). For example, the performance differences between the two separations were 4.7%, 5.3%, 11.0%, 9.5%, 3.9%, and 0.098 in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics, respectively, on the ACP740 dataset. On the ACPmain dataset, the differences reached 4.7%, 7.6%, 8.8%, 5.9%, 4.7%, and 0.094, respectively. The differences were 1.6%, 0.3%, 3.2%, 2.8%, 1.5%, and 0.031, respectively, on the Xiao dataset.

For comparison purposes, the TVI method was also tested on the 10 training and validation set separations generated by random sampling, and it performed better than the traditional training approach, with average improvement rates of 2.0%, 2.1%, 1.9%, 1.9%, 2.0%, and 4.6% in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics, respectively, on the ACP740 dataset (see Figure 2). On the ACPmain dataset, the average improvement rates were 2.0%, 2.1%, 1.9%, 1.9%, 2.0%, and 4.7%, respectively (see Figure 3). The average improvement rates reached 0.47%, 0.24%, 0.72%, 0.63%, 0.45%, and 0.98%, respectively, on the Xiao dataset

(see Figure 4). Moreover, the largest improvement rates in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics were 3.9%, 6.6%, 8.2%, 6.6%, 4.4%, and 9.1% on the ACP740 dataset; 5.0%, 6.6%, 5.0%, 4.3%, 5.4%, and 16.9% on the ACPmain dataset; and 1.6%, 0.3%, 3.2%, 2.8%, 1.5%, and 3.3% on the Xiao dataset, respectively.

Moreover, for the TVI training method, we calculated the performance differences between each of the two separations of the training and validation sets and found that

the differences obviously decreased (see Table S1). The results demonstrated that the TVI method effectively reduced the biases of the tested network models on different separations of the training and validation sets.

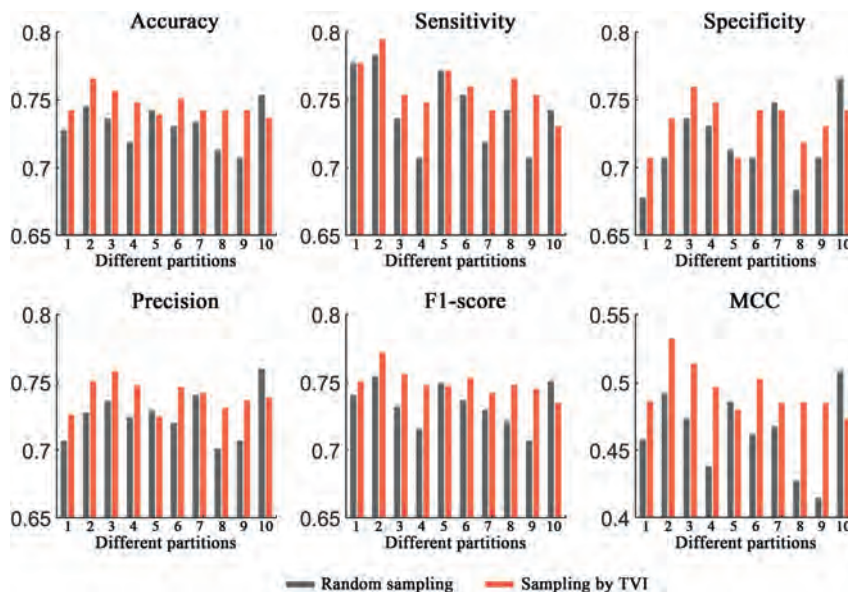
In addition, we evaluated the effectiveness of TVI on other ACP network models. Two models, ACP-DL and MHCNN, were employed to perform the test because they provided full codes that could be utilized to implement our TVI training method. After the evaluation was completed, the results demonstrated that the two network models still exhibited obvious biases on different separations of the training and validation sets, and the TVI training method still performed better than the traditional training method, suggesting its strong generalization ability (see Notes S1 and S2 and Figures S1–S4). The detailed test information of each model on different datasets can be seen at <https://github.com/wanyunzh/TriNet>.

Furthermore, to prevent the influence of the fixed test sets, a 5-fold cross-validation was also performed on the ACP740 dataset by using the TVI method. As shown in Figure 5, the TVI method consistently performed better than the traditional training approach, with improvement rates of 2.5%, 1.2%, 4.7%, 3.8%, 2.5%, and 6.0% in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics, respectively, on the ACP740 dataset, indicating that the TVI training method was not restricted to the test sets. Based on the above facts, we believe that the TVI method exhibits great potential for improving the performance of peptide predictions.

### Performance comparison with other existing models

#### Comparison with ACP predictors

In this section, we compared the performance of TriNet with that of several other state-of-the-art ACP predictors by conducting 5-fold cross-validations on the ACP740 dataset and independent tests on the ACPmain and ACPalternate datasets. On the ACP740 dataset, we compared TriNet with ACP-DL, MHCNN, iACP-DRLF, CL-ACP, and DeepACPPred.<sup>23–26,30</sup> On the ACPmain and ACPalternate datasets, we compared it with



**Figure 3. Performance comparison between the traditional training approach and the TVI method on the ACPmain dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

ACP-DL, MHCNN, iACP-DRLF, and AntiCP 2.0.<sup>23,24,26,31</sup> In these ACP prediction approaches, to our knowledge, a validation set is not established during model training when an independent test is performed. Harrington<sup>33</sup> indicated that a single split of the training and test sets can result in an inaccurate evaluation of the tested model's performance. Therefore, we randomly selected 20% of the peptides from the ACPmain and ACPalternate training datasets to form the validation sets. For a fair comparison, all the compared methods were retrained by using the same training and validation sets on the two datasets and then tested on the independent test sets.

After comparison, the results showed that TriNet performed the best among all the compared methods on all three datasets. In detail, the improvement rates achieved by TriNet over the other compared methods were 3.2%–8.6%, 1.9%–6.9%, 3.2%–21.5%, 3.0%–12.0%, 3.2%–6.9%, and 6.9%–22.9% on the ACP740 dataset (see Figure 6) in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics, respectively). The improvement rates in comparison with other methods were 3.2%–12.0%, 5.4%–17.2%, 1.9%–9.5%, 3.6%–13.2%, and 9.8%–44.7% on the ACPmain independent dataset in terms of accuracy, sensitivity, precision, F1 score, and MCC metrics, respectively (see Figure 7), and 1.7%–9.0%, 2.9%–9.3%, 2.0%–9.3%, and 3.3%–21.4% on the ACPalternate independent dataset in terms of accuracy, sensitivity, F1 score, and MCC metrics, respectively (see Figure 8). It should be noted that, although the specificity of AntiCP-2.0 is slightly higher than that of TriNet on the ACPmain independent dataset, its other indicators are lower than those of TriNet. Similarly, on the ACPalternate independent dataset, the specificity and precision of iACP-DRLF are slightly higher than those of TriNet, while its other indicators are lower than those of TriNet. Therefore, TriNet demonstrated the best overall performance on all three ACP datasets.

#### Comparison with AMP predictors

In addition to ACP predictors, we compared TriNet with AMP prediction tools, including DNN, APIN, ACEP, CAMP-RF,

CAMP-SVM, and CAMP-ANN,<sup>20–22,32</sup> by testing them on three AMP datasets. After comparison, the results showed that TriNet performed better than all the compared methods on the three datasets. In detail, the improvement rates achieved by TriNet over the other compared methods were 2.8%–9.6%, 0.78%–5.7%, 3.8%–16.9%, 3.5%–13.3%, 2.7%–9.0%, and 6.0%–21.7% on the Xiao independent dataset (see Figure 9) in terms of the accuracy, sensitivity, specificity, precision, F1 score, and MCC metrics, respectively, and 1.0%–20.8%, 3.9%–10.5%, 0.13%–26.9%, 0.23%–26.9%, 1.1%–18.7%, and 2.2%–57.2% on the AMPlify dataset (see Figure 10), respectively. On the DAMP dataset (see Figure 11), the improvement rates in terms of accuracy, specificity, precision, F1 score, and MCC were 1.4%–10.7%, 1.7%–18.4%, 1.8%–15.8%, 1.3%–10.8%, and 3.1%–26.5%. Although the sensitivity of TriNet is lower than that of CAMP-RF, its other indicators are much higher, demonstrating that the best overall performance is achieved using TriNet (see Figure 11).

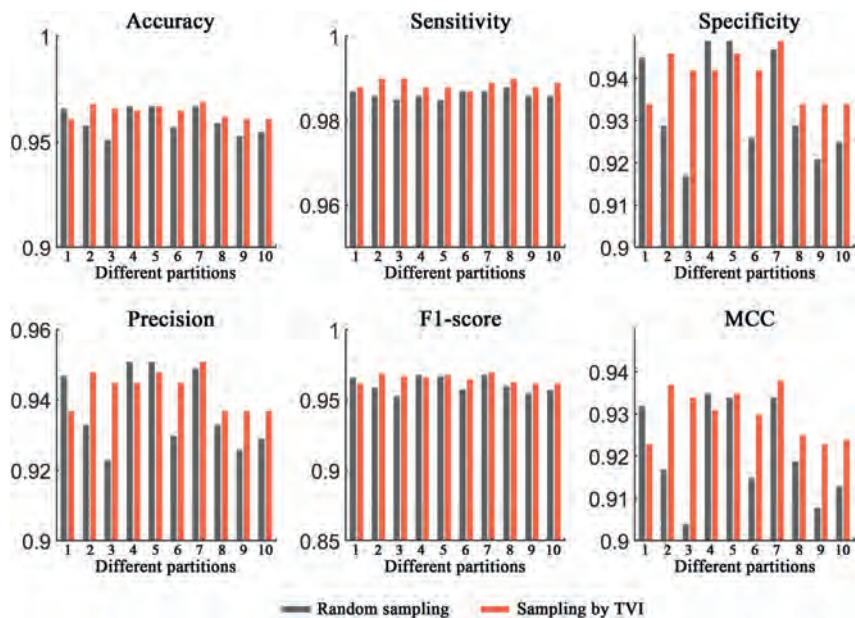
#### Evaluation of the effectiveness of the extracted features and the network structures

In this paper, we carried out multiple tests on the three datasets, ACP740, ACPmain, and Xiao to verify the effectiveness of our feature extraction methods as well as the superiority of the network structures.

##### Effectiveness of the extracted features

In this section, we first demonstrated the advantages of the improved DCGR method over the original DCGR approach in terms of extracting the sequence serial fingerprint features. Then, we verified the importance of combining all three features. Finally, we demonstrated the extensive adaptability and effectiveness of the three features.

To demonstrate the advantages of the improved DCGR method, we compared its performance with that of the original DCGR approach on all three datasets, and the results showed that the improved method performed better than the original techniques in terms of all the metrics on the three datasets (see Table S2). Then, we attempted to verify the importance of combining all three features by removing each feature individually, and the results showed that the loss of any of the three features resulted in performance degradation on all three datasets (see Table S3). In addition, in comparison with the physicochemical property feature, the removal of the serial fingerprint or sequence evolution feature caused a more serious performance decline. Finally, to demonstrate the extensive adaptability and effectiveness of the three features, we replaced the neural



**Figure 4. Performance comparison between the traditional training approach and the TVI method on the Xiao dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

(see Table S9). The reason for this may be that a single head is able to make the network cover the most effective information concerning the distribution of the physicochemical properties. As a consequence, the use of multiple heads makes the model fail to capture the differences among the heads, and finally, the multi-head models become more complex and ineffective.

## DISCUSSION

In this study, we introduced TriNet, a trifusion neural network for ACP or AMP prediction. After evaluating the performance of TriNet and comparing it with other leading prediction methods on multiple challenging datasets, we found that TriNet demonstrated much higher accuracy in terms of predicting both ACPs and AMPs than all the compared methods under commonly used criteria. The superiority of TriNet may be attributed to the following method innovations.

First, we proposed that a prediction method for ACPs and AMPs should effectively fuse multiple pieces of information, based on which the TriNet framework was designed. Second, in addition to the frequently used sequence evolution feature, we introduced another two features, the serial fingerprint and physicochemical property features, which appropriately characterize the global sequence information and the distributions of the physicochemical properties of peptides. The test results demonstrated the extensive adaptability and effectiveness of the proposed features. Third, based on the properties of the three features, we specifically designed three network structures, which appropriately processed each of the features and then effectively fused them for the final predictions. Fourth, we developed a neural network training approach called TVI, which was able to generate more appropriately separated training and validation sets based on the biases of a network model. In addition, we provided the learning curves of all six datasets (see Figures S6–S11) to demonstrate the degree of overfitting, and it was shown that there is no obvious overfitting phenomenon on any of these datasets.

In supervised deep-learning fields, setting both the validation and the test sets is of great importance for evaluating the generalization ability of a network model according to the predictive power of blind test sets. However, as we know, in the field of ACP prediction, many models have only training and test sets, which clearly leads to information leakage from the test set and an inaccurate evaluation of the model's performance. In contrast, we set the validation sets for both 5-fold cross-validation and independent testing.

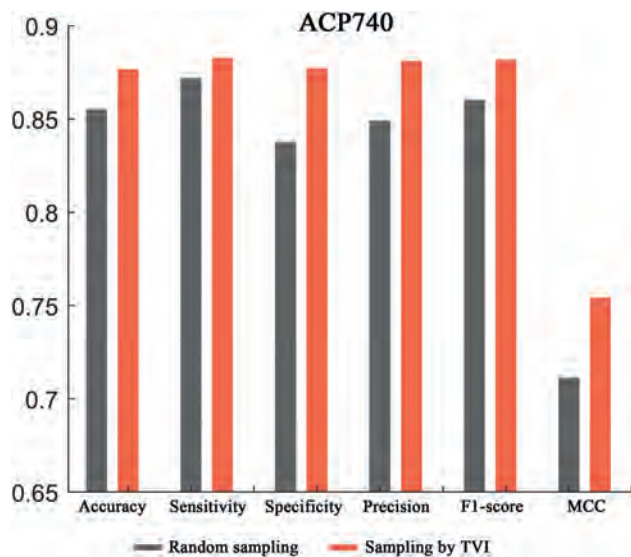
network with the XGBoost<sup>34</sup> algorithm, which is a popular traditional machine learning technique, for retraining the samples on all three datasets. In detail, the three feature matrices obtained from DCGR, PSSM, and physicochemical property embedding (PCPE) were first flattened and then concatenated to generate a feature vector for XGBoost. The testing results (see Tables S4–S6) showed that XGBoost achieved higher performance than many of the compared models on all three datasets, demonstrating the extensive adaptability and effectiveness of the features extracted in this study.

### Effectiveness of the network structures

Regarding the CNN-CAM mechanism for processing the serial fingerprint feature, we replaced the  $1 \times 8$  kernel with three frequently used square kernels with sizes of 2, 3, and 5, and the results showed that our property-based  $1 \times 8$  kernel performed the best (see Table S7), mainly because such a kernel size made the network learn the shared weights based on each physicochemical property. The CAM contains two pooling strategies: average pooling and maximum pooling. We first replaced this module with the squeeze-and-excitation network (SENet),<sup>35</sup> which uses only global average pooling, and the prediction performance obviously decreased on all three datasets (see Table S8). As shown in previous studies,<sup>36</sup> maximum pooling compensates for the global information gained from average pooling by reflecting the salient part of each channel. Then, we replaced the CAM with another popular convolutional block attention module (CBAM)<sup>36</sup> that has a spatial attention module (SAM) after its CAM, and we found that the prediction performance still declined on most datasets (see Table S8). Since the SAM mainly focuses on the importance of the spatial features, the serial fingerprint feature extracted by DCGR had no spatial or positional properties.

For the encoder module, we tested different numbers of heads in the self-attention module. In detail, we set 1, 2, and 4 heads and compared the resulting performances. The results showed that the single-head self-attention mechanism performed better than the multihead self-attention mechanism





**Figure 5. Comparison between the traditional training approach and the TVI method conducted via 5-fold cross-validation on the ACP740 dataset**

The comparison was performed under six different evaluation metrics: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

Despite the obvious advantages of TriNet, we still have a long way to go to completely solve the ACP/AMP prediction problem, and further improvements can still be made on TriNet in the future. For example, the current model is not an end-to-end model. Thus, it still takes some time to calculate the corresponding features. Therefore, the inference time of TriNet may be longer than that of end-to-end frameworks. In addition, we note that the current version of TVI may perform slightly worse than traditional training methods in some cases, and more attention should be given to the following issues. (1) How can the starting epoch for interaction among the samples in the training and validation sets be determined? (2) How can the number of interacting samples between the two sets be determined? (3) How can the interaction termination epoch be determined? Moreover, the issue of determining whether interactions are required for the given separation of the training and validation sets still needs to be further investigated. The future version of TriNet will attempt to solve these problems and make further improvements.

The results of the evaluations showed that our method could clearly distinguish between ACPs/AMPs and non-ACPs/AMPs, and the potential of TriNet for identifying ACPs/AMPs will help researchers develop small-peptide drugs targeting cancer cells or other pathogens, such as bacteria, fungi, and viruses. In addition, the TVI training method may become the next trend for training different neural networks in other areas.

## EXPERIMENTAL PROCEDURES

### Resource availability

#### Lead contact

The lead contact for questions about this paper is Juntao Liu, who can be reached at [juntaosdu@126.com](mailto:juntaosdu@126.com).

### Materials availability

No unique materials were generated from this study.

### Data and code availability

The data that support the findings of this study are available from the lead contact upon reasonable request. The authors declare that all other data supporting the findings of this study are available within the paper and its supplemental information files. TriNet is deployed on our web server: <http://liulab.top/TriNet/server>. All original code has been deposited at Zenodo under <https://doi.org/10.5281/zenodo.7556870> and is publicly available as of the date of publication.

## Methodology

### Dataset preparation

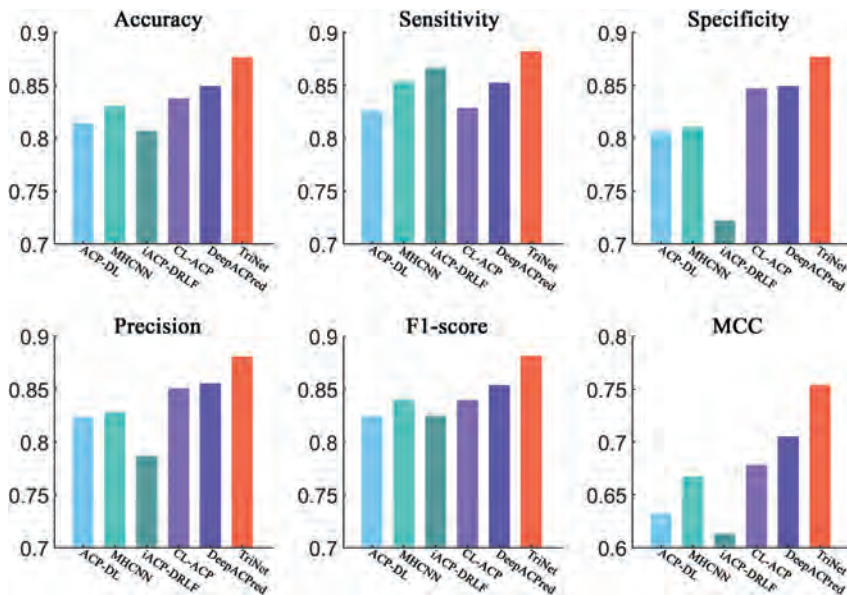
In this study, six datasets were collected to test the prediction performance of TriNet, including three ACP datasets (ACP740 dataset, ACPmain dataset, and ACPalternate dataset) for ACP prediction and three AMP datasets (Xiao dataset, DAMP dataset, and AMPlify dataset) for AMP prediction. ACP740 was introduced by Yi et al.<sup>23</sup>; it contains 376 experimentally validated ACPs and 364 AMPs without anticancer activity, and the sequence similarity between each pair of peptides is no greater than 90%. ACPmain and ACPalternate were introduced from Agrawal et al.,<sup>31</sup> and each dataset contains two subsets. The first subset of ACPmain, which includes 689 experimentally validated ACPs and 689 non-ACPs, was separated into two parts for training and validation with a 4:1 ratio. The second subset of ACPmain, which includes 172 experimentally validated ACPs and 172 non-ACPs, was used as the independent test set. The first subset of ACPalternate, which includes 776 experimentally validated ACPs and 776 non-ACPs, was also separated into two parts for training and validation with a 4:1 ratio. The second subset of ACPmain, which includes 194 experimentally validated ACPs and 194 non-ACPs, was used as the independent test set.

For the AMP datasets, Xiao's benchmark training dataset<sup>37</sup> comprises 1,388 AMPs and 1,440 non-AMPs, and the corresponding independent test set comprises 920 AMPs and 920 non-AMPs. However, the dataset from Xiao<sup>37</sup> has a major difference in length distribution between AMP and non-AMP sequences. We followed the same method as Veltri et al.<sup>20</sup> and randomly adjusted the lengths of non-AMP sequences to more closely resemble AMP sequences to avoid learning the length differences. The AMP and non-AMP sequence length distributions of the original dataset and our readjusted dataset are provided in Figures S12 and S13. Then, Xiao's training dataset was divided into two parts for training and validation with a 4:1 ratio. DAMP was introduced by Veltri et al.,<sup>20</sup> and the 3,556 peptide sequences (1,778 AMPs and 1,778 non-AMPs) with a similarity of no more than 40% were divided into three parts: 1,424 for training, 708 for validation, and 1,424 for testing. AMPlify was introduced by Li et al.,<sup>38</sup> and the non-AMP sequences in the dataset were also adjusted to match the length distributions of the AMP sequences. The training dataset comprising 3,338 AMPs and 3,338 non-AMPs was also divided into two parts for training and validation with a 4:1 ratio, and the independent test set comprises 835 AMPs and 835 non-AMPs.

### Overview of the TriNet framework

The TriNet pipeline was designed to predict ACPs and AMPs based solely on the given peptide sequences. In this study, we assumed that whether a peptide was an ACP or AMP could be predicted by effectively combining three kinds of features representing the serial fingerprints, sequence evolutions, and physicochemical properties of peptide sequences. Therefore, the main architecture of the TriNet pipeline comprises three parallel components, a CNN-CAM, a Bi-LSTM network, and an encoder module, for processing and fusing the above three features (Figure 1). By using batch normalization, 400-, 256-, and 50-dimensional feature vectors were obtained as the outputs of each branch. These feature vectors were then concatenated and passed through dropout and dense layers to generate the final prediction results. The final dense layer employs a sigmoid function generating a score in [0,1] to determine that the peptide is an ACP/AMP if the score is no smaller than 0.5 and a non-ACP/AMP otherwise.

Moreover, to obtain better training results than those of the traditional training method, which randomly separated the training and validation datasets, a training approach termed TVI was designed to reparate the training and validation sets based on the structure of the neural network, which was



**Figure 6. Comparison of TriNet with existing models on the ACP740 dataset using 5-fold cross-validation**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

achieved through iterative interaction of the samples in the training and validation datasets. In the following sections, we introduce each part of the TriNet method in detail.

#### Extraction of peptide sequence features

Given a peptide sequence, three kinds of features reflecting the information of the serial fingerprints, sequence evolutions, and physicochemical properties of peptide sequences were extracted as follows.

**Extracting the serial fingerprint features of sequences.** DCGR<sup>39</sup> is a protein sequence feature extraction method based on chaotic game representation (CGR)<sup>39,40</sup> that attempts to capture the global characteristics of a protein sequence; therefore, the extracted features can effectively reflect the serial fingerprint information of the given peptide sequences (see Note S3 for the method details). The original DCGR method obtained distance matrices only in four quadrants, and the information between the points that crossed quadrants was therefore lost. To recover this lost information, we improved the DCGR method by rotating the coordinate axis by 45° to obtain another four distance matrices (see Figure S5). Then, for each CGR curve, eight distance matrices,  $A_{11}, A_{12}, \dots, A_{18}$ , could be calculated, and the final  $158 \times 8$  feature matrix  $M_{DCGR}$  for each peptide could be expressed as:

$$d_i = [\rho(A_{11}), \rho(A_{12}), \dots, \rho(A_{18})]^T, \quad (\text{Equation 1})$$

$$M_{DCGR} = [d_1, d_2, \dots, d_i, \dots, d_{158}]^T, \quad (\text{Equation 2})$$

where  $\rho(A_{ij})$  denotes the leading eigenvalues of the distance matrix  $A_{ij}$  and 158 represents the 158 physicochemical properties selected from the AAindex.

**Extracting sequence evolution features.** The PSSM is frequently applied to detect distant homologs using iterations.<sup>41,42</sup> An element  $(i, j)$  in the PSSM is proportional to the probability of the residue at position  $i$  being replaced by amino acid  $j$ , reflecting the evolutionary information of peptide sequences. The PSI-BLAST<sup>43</sup> tool was employed to obtain an  $L \times 20$  feature matrix  $M_{PSSM}$  for each peptide.

**Extracting sequence physicochemical property features.** PCPE is capable of reflecting the distributions of physicochemical properties in peptide sequences. Regarding the choice of physicochemical properties, traditional methods usually select specific physicochemical properties directly and therefore may result in the chosen physicochemical properties exhibiting redundancy or low quality. In contrast, we first employed the method proposed by Saha et al.<sup>44</sup> to group the 556 physicochemical properties into eight clusters and extracted the most representative property in each cluster to obtain more comprehensive physicochemical properties while avoiding redundancies. Then, by using PCPE, each amino acid was encoded into an eight-

dimensional vector, and an  $L \times 8$  feature matrix  $M_{PCPE}$  was finally constructed for each peptide.

To make the feature matrices  $M_{PSSM}$  and  $M_{PCPE}$  of all the peptides with different lengths have the same dimensions, we set the sequence length  $L$  to 50 and used zero-padding for peptides whose lengths were less than 50.

#### Processing of the serial fingerprint features via the CNN-CAM module

The feature matrix  $M_{DCGR}$  obtained from DCGR was reshaped into a three-dimensional tensor and fed into an improved CNN (see Figure 1B), which is capable of capturing important features through local connectivity and weight sharing. Traditional CNNs usually apply square kernels to learn to convolve feature matrices. However, each row of

the feature matrix  $M_{DCGR}$  denotes one of the 158 physicochemical properties, and the columns represent the eight features extracted from one CGR curve. Therefore, more appropriate kernels of size  $1 \times 8$  (instead of the frequently used square kernels) were applied by TriNet to learn the shared weights for each of the 158 properties. The number of filters was set to 16 in this study.

The CNN effectively captured the local information from each physicochemical property, based on which the CAM<sup>36</sup> was then employed to obtain the global information by emphasizing important features from all 158 properties. The CAM model is able to emphasize more valuable features by assigning larger weights. In detail, after calculating the three-dimensional feature map  $M'_{DCGR} = f_{conv}(M_{DCGR})$  from the convolutional layer, each channel  $C_i$  of  $M'_{DCGR}$  was assigned a channel weight  $CAM_i$  according to the classification importance of this channel. First, global average and maximum pooling were performed on the feature map  $M'_{DCGR}$ , followed by a shared multilayer perceptron (MLP) comprising two dense layers. The whole process of the CAM can be formulated as follows:

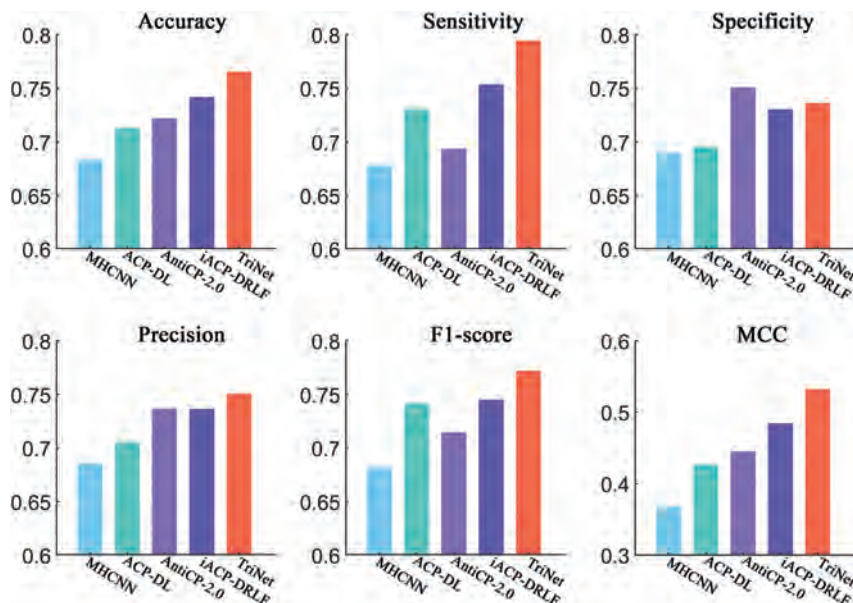
$$CAM(M'_{DCGR}) = \sigma\{MLP[AvgPool(M'_{DCGR})] + MLP[MaxPool(M'_{DCGR})]\} \\ = \sigma\{W_1[W_0(M'_{avg})] + W_1[W_0(M'_{max})]\}, \quad (\text{Equation 3})$$

where  $\sigma$  denotes the sigmoid function,  $M'_{DCGR} \in R^{158 \times 1 \times 16}$  and  $M'_{avg}, M'_{max} \in R^{1 \times 1 \times 16}$  are two matrices that calculate the average and maximum pooling, respectively, and  $W_0 \in R^{2 \times 16}$  (with the rectified linear unit [ReLU] activation function) and  $W_1 \in R^{16 \times 2}$  represent the weight matrices of the shared MLP.

The channel weights were then assigned to the corresponding channels of the feature map  $M'_{DCGR}$  for element-wise multiplication, and the weight-assigned feature map  $M''_{DCGR} \in R^{158 \times 1 \times 16}$  was generated. Then,  $M''_{DCGR}$  was flattened and passed through a dense layer and transformed into the final DCGR feature vector  $F_{DCGR}$  with 400 dimensions.

#### Processing of the sequence evolution features via the Bi-LSTM layer

As shown in Figure 1C, the feature matrix  $M_{PSSM}$  obtained from the PSSM was fed into the Bi-LSTM layer. Different from the traditional RNN, the LSTM network<sup>45</sup> is able to learn and capture both the long- and the short-term dependencies among the amino acids of a peptide sequence. Moreover, studies have shown that certain types of residues are usually favored at the N terminus and C terminus of ACPs and AMPs, which play crucial roles in identifying ACPs and AMPs.<sup>15,46</sup> Therefore, by analyzing the peptide sequences in the forward and backward directions, Bi-LSTM is capable of obtaining information from the C terminus and N terminus for peptides with lengths of no more than 50 amino acids at the same timestep. The calculations of the forward LSTM can be summarized as follows:



**Figure 7. Comparison of TriNet with existing models on the ACPmain independent dataset**  
Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

$$f_t = \sigma(W_{hf} h_{t-1} + W_{xt} x_t + b_f), \quad (\text{Equation 4})$$

$$i_t = \sigma(W_{hi} h_{t-1} + W_{xi} x_t + b_i), \quad (\text{Equation 5})$$

$$\tilde{c}_t = \tanh(W_{hc} h_{t-1} + W_{xc} x_t + b_c), \quad (\text{Equation 6})$$

$$O_t = \sigma(W_{ho} h_{t-1} + W_{xo} x_t + b_o), \quad (\text{Equation 7})$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t, \quad (\text{Equation 8})$$

$$h_t = O_t \otimes \tanh(c_{t-1}), \quad (\text{Equation 9})$$

where  $t = 1, 2, \dots, 50$  represents the order of 50 amino acids of a peptide sequence;  $W_{hf}$ ,  $W_{hi}$ ,  $W_{hc}$ ,  $W_{ho}$ ,  $W_{xf}$ ,  $W_{xi}$ ,  $W_{xc}$ , and  $W_{xo}$  are weight matrices;  $b_f$ ,  $b_i$ ,  $b_c$ , and  $b_o$  are bias vectors;  $f_t$  is the forget gate;  $i_t$  is the input gate;  $o_t$  is the

output gate;  $x_t$  is the current input;  $c_{t-1}$  is the previous cell state;  $c_t$  is the current cell state;  $\tilde{c}_t$  is the value added to the cell state;  $h_{t-1}$  and  $h_t$  are the previous and current hidden states, respectively; and  $\otimes$  represents the elementwise multiplication operations.

The backward LSTM works in the same way as the forward LSTM with the calculated current hidden state being  $h'_t$ . The final PSSM feature vector is then formulated as  $F_{PSSM} = [h_t, h'_t]$  of 256 dimensions, with  $t$  being the last time step.

#### Processing the physicochemical property features via the encoder module

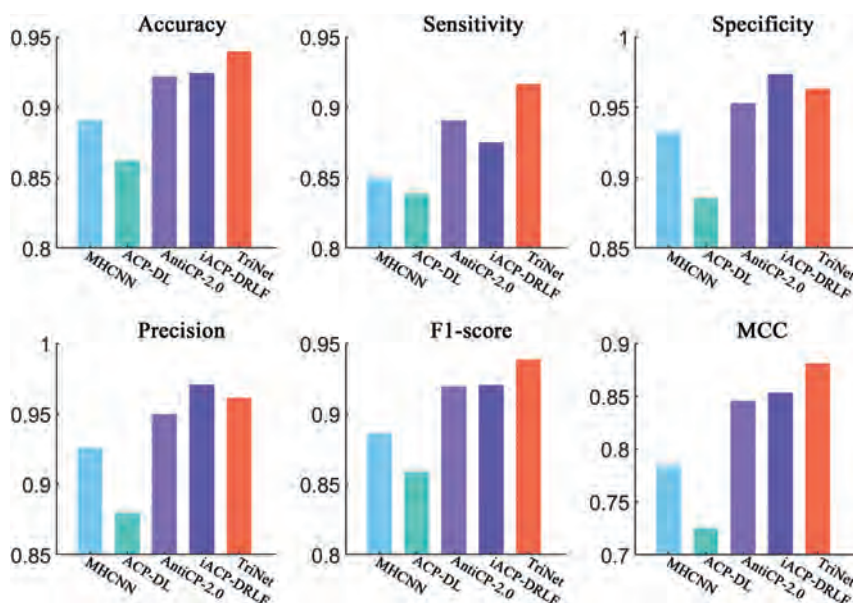
The feature matrix  $M_{PCPE}$  obtained from PCPE was fed into the encoder block, with each row representing an eight-dimensional embedding vector (see Figure 1D). The encoder block was designed

based on the encoder of a transformer,<sup>47</sup> which contains multihead self-attention mechanisms, a feedforward network, and skip connections followed by layer normalization. The main part of the transformer is multihead self-attention, which is able to calculate the dependencies between amino acid residues despite the long distances between them, hence efficiently capturing the dependency information of the physicochemical properties of specific peptides. In this paper, single-head self-attention was employed, and its calculation process is summarized as follows:

$$q_i = W_q p_i, k_i = W_k p_i, v_i = W_v p_i, i = 1, 2, \dots, L, \quad (\text{Equation 10})$$

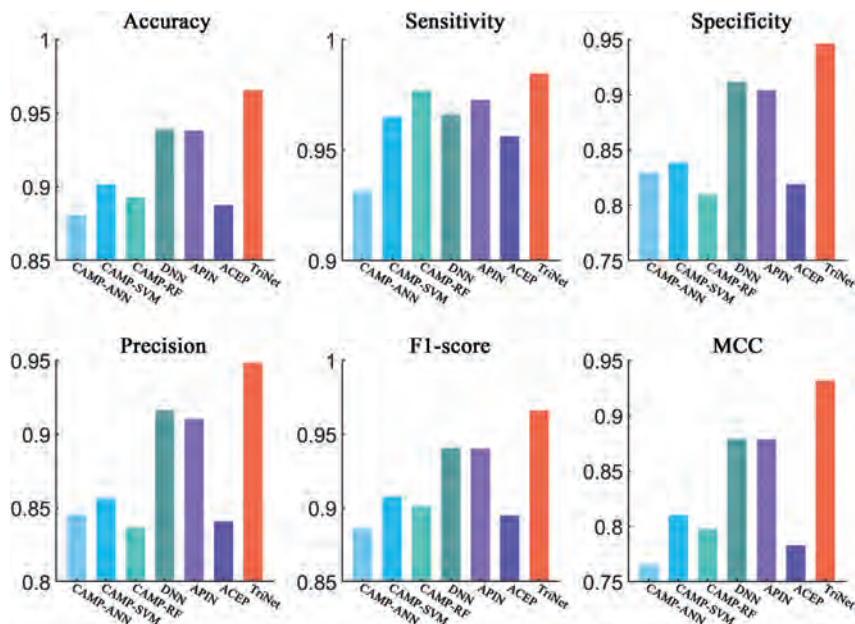
$$Q = [q_1, q_2, \dots, q_L]^T, K = [k_1, k_2, \dots, k_L]^T, V = [v_1, v_2, \dots, v_L]^T, \quad (\text{Equation 11})$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (\text{Equation 12})$$



**Figure 8. Comparison of TriNet with existing models on the ACPalternate independent dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.



**Figure 9. Comparison of TriNet with existing models on Xiao's independent dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

Then, a feature matrix  $M'_{PCPE} = [p_1, p_2 \dots p_L]^T$  obtained by adding the positional encoding information was constructed, with  $p_i$  denoting the feature vector of the  $i$ -th residue and  $L = 50$  representing the sequence length. Passing through the encoder module, average pooling was applied, and the final 50-dimensional PCPE feature vector  $F_{PCPE}$  was calculated for each peptide.

**Network training by iterative interaction between the training and validation sets**

After constructing a neural network, traditional training methods usually randomly separate the training and validation sets and then train the model on the training set and validate it on the validation set. In fact, neural networks may show great biases on different separations, and therefore, different separations of the training and validation sets may

where  $A$  is the attention score matrix;  $q_i$ ,  $k_i$ , and  $v_i$  are query, key, and value vectors, respectively;  $d_k$  is their dimensionality; and  $W_q$ ,  $W_k$ , and  $W_v \in R^{d_k \times d_p}$  are the corresponding weight matrices.

Furthermore, since the order of the residues plays a crucial role in a peptide sequence, positional encoding, using the sine and cosine to reflect the distribution of the physicochemical properties in a peptide sequence, was applied in this study as follows:

$$PE(pos, 2i) = \sin(pos/10000^{2i/d_p}), \quad (\text{Equation 13})$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{2i/d_p}), \quad (\text{Equation 14})$$

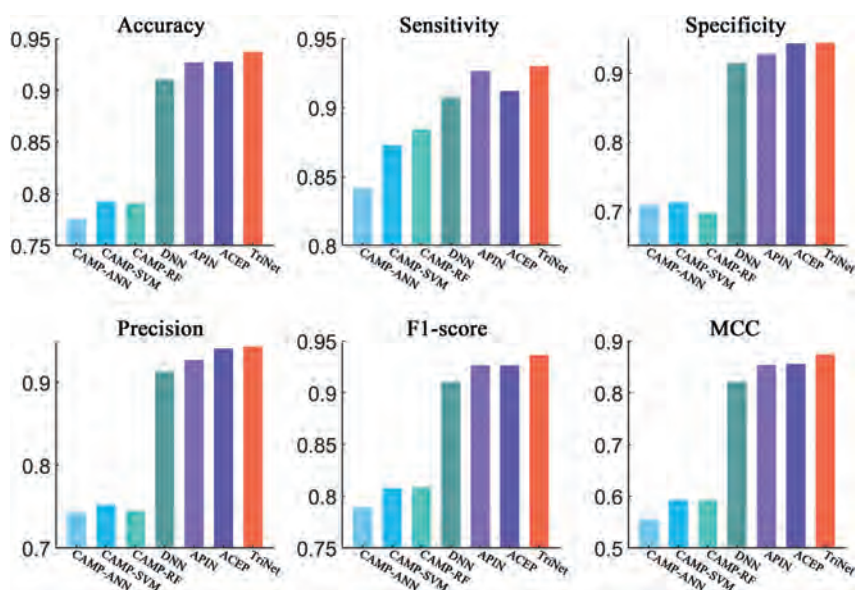
where  $pos$  represents the positions of the amino acids in the sequence,  $2i$  and  $2i + 1$  denote the even and odd element sites in the embedding vectors, respectively, and  $d_p = 8$  is the dimensionality of the embedding vectors.

largely influence the training of the network model and hence the performance achieved on the testing set. To construct more appropriate training and validation sets by considering the biases of a specific neural network, a method termed TVI was proposed by iteratively interacting the samples in the training and validation sets as follows.

Step 1. Randomly separate a training set  $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  and a validation set  $V = \{(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_m, y'_m)\}$ , where  $x_i$  and  $x'_i$  are the feature vectors of the samples in the training and validation sets, respectively, and  $y_i$  and  $y'_i \in \{0, 1\}$  are the sample labels. Train and validate the constructed network model on the two sets for  $N$  epochs.

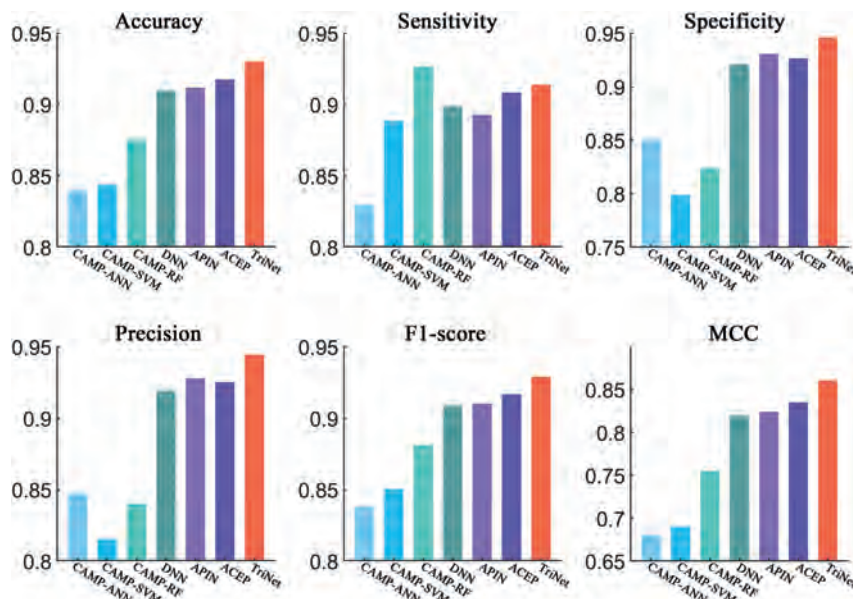
Step 2. Search for the samples in  $V$  that are erroneously classified more than five times in the last 10 epochs, termed  $V' = \{(x'_{m_1}, y'_{m_1}), (x'_{m_2}, y'_{m_2}), \dots, (x'_{m_k}, y'_{m_k})\}$ , and search for the samples in  $T$  that are correctly classified in each of the last 10 epochs, termed  $T' = \{(x_{n_1}, y_{n_1}), (x_{n_2}, y_{n_2}), \dots, (x_{n_l}, y_{n_l})\}$ .

Step 3. Randomly select  $[k/2]$  samples from  $V'$ , termed  $V_{change}$ , and  $[k/2]$  samples from  $T'$ , termed  $T_{change}$  (if  $[k/2]$  is larger than  $l$ , then randomly select  $l$  samples



**Figure 10. Comparison of TriNet with existing models on the AMPlify dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.



**Figure 11. Comparison of TriNet with existing models on the DAMP dataset**

Six different evaluation metrics are shown: accuracy, sensitivity, specificity, precision, F1 score, and MCC.

from  $V'$  and  $T'$ ). Then, construct a training set  $T_{new}$  and a validation set  $V_{new}$  by exchanging the samples of  $T$  and  $V$  that are contained in  $T_{change}$  and  $V_{change}$ .

Step 4. Retrain the network model on the two sets  $T_{new}$  and  $V_{new}$ , repeat step 3 and step 4  $M$  ( $M$  was set to 2 in this study) times, and obtain the final training and validation sets  $T_{final}$  and  $V_{final}$ . Then, reinitialize the neural network and perform training and validation on  $T_{final}$  and  $V_{final}$ .

#### Evaluation metrics and methods

In this study, the widely used accuracy (Acc), sensitivity (Sens), specificity (Spec), precision (Prec), F1 score, and MCC criteria were applied to evaluate the performance of the models (see Note 4 for the definitions of these criteria). To evaluate the effectiveness of the models, 5-fold cross-validation and independent testing were employed on multiple datasets. For the 5-fold cross-validation, we randomly divided all the samples into five sets of equal size, among which four were used for training and validation (the training-validation ratio was 4:1), and the remaining set was used for testing. This process was repeated five times in such a way that each of the five sets was used once for testing, and the final performance was obtained by averaging the performance achieved across all five sets.

#### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.patter.2023.100702>.

#### ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China with code 2020YFA0712400 and the National Natural Science Foundation of China with codes 61801265 and 62272268. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

#### AUTHOR CONTRIBUTIONS

Conceptualization, J.L., W.Z., and Y. Liu; methodology, W.Z., Y. Liu, and Y. Li; formal analysis, W.Z., Y. Liu, and C.M.; resources, W.Z., Y. Liu, S.K., W.W., and B.D.; writing – original draft, W.Z. and Y. Liu; writing – review & editing, J.L., W.Z., Y. Liu, and X.G.; software, W.Z., Y. Li, and Y. Liu; visualization, J.H. and W.Z.; supervision, J.L.

#### DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: December 9, 2022

Revised: December 20, 2022

Accepted: February 3, 2023

Published: February 28, 2023

#### REFERENCES

- Roca, I., Akova, M., Baquero, F., Carlet, J., Cavaleri, M., Coenen, S., Cohen, J., Findlay, D., Gyssens, I., Heuer, O.E., et al. (2015). The global threat of antimicrobial resistance: science for intervention. *New Microbes New Infect.* 6, 22–29. <https://doi.org/10.1016/j.nmni.2015.02.007>.
- Hwang, A.Y., and Gums, J.G. (2016). The emergence and evolution of antimicrobial resistance: impact on a global scale. *Bioorg. Med. Chem.* 24, 6440–6445. <https://doi.org/10.1016/j.bmc.2016.04.027>.
- Nuti, R., Goud, N.S., Saraswati, A.P., Alvala, R., and Alvala, M. (2017). Antimicrobial peptides: a promising therapeutic strategy in tackling antimicrobial resistance. *Curr. Med. Chem.* 24, 4303–4314. <https://doi.org/10.2174/0929867324666170815102441>.
- Mookherjee, N., Anderson, M.A., Haagsman, H.P., and Davidson, D.J. (2020). Antimicrobial host defence peptides: functions and clinical potential. *Nat. Rev. Drug Discov.* 19, 311–332. <https://doi.org/10.1038/s41573-019-0058-8>.
- Bhandari, V., and Suresh, A. (2022). Next-Generation approaches needed to tackle antimicrobial resistance for the development of novel therapies against the deadly pathogens. *Front. Pharmacol.* 13. <https://doi.org/10.3389/fphar.2022.838092>.
- Reddy, K., Yedery, R., and Aranha, C. (2004). Antimicrobial peptides: pre-mises and promises. *Int. J. Antimicrob. Agents* 24, 536–547. <https://doi.org/10.1016/j.ijantimicag.2004.09.005>.
- Greco, I., Molchanova, N., Holmedal, E., Jenssen, H., Hummel, B.D., Watts, J.L., Håkansson, J., Hansen, P.R., and Svenson, J. (2020). Correlation between hemolytic activity, cytotoxicity and systemic in vivo toxicity of synthetic antimicrobial peptides. *Sci. Rep.* 10, 13206. <https://doi.org/10.1038/s41598-020-69995-9>.
- Huan, Y., Kong, Q., Mou, H., and Yi, H. (2020). Antimicrobial peptides: classification, design, application and research progress in multiple fields. *Front. Microbiol.* 2559, 582779. <https://doi.org/10.3389/fmicb.2020.582779>.

9. Zhong, W., Zhong, B., Zhang, H., Chen, Z., and Chen, Y. (2019). Identification of anti-cancer peptides based on multi-classifier system. *Comb. Chem. High Throughput Screen.* 22, 694–704. <https://doi.org/10.2174/1386207322666191203141102>.
10. Ng, C.X., Le, C.F., Tor, Y.S., and Lee, S.H. (2021). Hybrid anticancer peptides DN1 and DN4 exert selective cytotoxicity against hepatocellular carcinoma cells by inducing both intrinsic and extrinsic apoptotic pathways. *Int. J. Pept. Res. Ther.* 27, 2757–2775. <https://doi.org/10.1007/s10989-021-10288-8>.
11. Arpornsuwan, T., Sriwai, W., Jaresitthikunchai, J., Phaonakrop, N., Sritanaudomchai, H., and Roytrakul, S. (2014). Anticancer activities of antimicrobial BmKn2 peptides against oral and colon cancer cells. *Int. J. Pept. Res. Ther.* 20, 501–509. <https://doi.org/10.1007/s10989-014-9417-9>.
12. Chen, L., Jia, L., Zhang, Q., Zhou, X., Liu, Z., Li, B., Zhu, Z., Wang, F., Yu, C., Zhang, Q., et al. (2017). A novel antimicrobial peptide against dental caries-associated bacteria. *Anaerobe* 47, 165–172. <https://doi.org/10.1016/j.anaerobe.2017.05.016>.
13. Björn, C., Noppa, L., Näslund Salomonsson, E., Johansson, A.-L., Nilsson, E., Mahlapuu, M., and Håkansson, J. (2015). Efficacy and safety profile of the novel antimicrobial peptide PXL150 in a mouse model of infected burn wounds. *Int. J. Antimicrob. Agents* 45, 519–524. <https://doi.org/10.1016/j.ijantimicag.2014.12.015>.
14. Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. <https://doi.org/10.1093/bioinformatics/bty451>.
15. Gautam, A., Chaudhary, K., Kumar, R., Sharma, A., Kapoor, P., Tyagi, A., and Raghava, G.P. (2013). In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.* 11, 1–12. <https://doi.org/10.1186/1479-5876-11-74>.
16. Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. <https://doi.org/10.1016/j.jtbi.2010.12.024>.
17. Wei, L., Tang, J., and Zou, Q. (2017). SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genom.* 18, 742. <https://doi.org/10.1186/s12864-017-4128-1>.
18. Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.H. (1999). Recognition of a protein fold in the context of the SCOP classification. *Proteins* 35, 401–407. [https://doi.org/10.1002/\(sici\)1097-0134\(19990601\)35:4<401::Aid-prot3>3.0.Co;2-k](https://doi.org/10.1002/(sici)1097-0134(19990601)35:4<401::Aid-prot3>3.0.Co;2-k).
19. Nasiri, F., Atanaki, F.F., Behrouzi, S., Kavousi, K., and Bagheri, M. (2021). Cpacpp: in silico cell-penetrating anticancer peptide prediction using a novel bioinformatics framework. *ACS Omega* 6, 19846–19859. <https://doi.org/10.1021/acsomega.1c02569>.
20. Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 34, 2740–2747. <https://doi.org/10.1093/bioinformatics/bty179>.
21. Su, X., Xu, J., Yin, Y., Quan, X., and Zhang, H. (2019). Antimicrobial peptide identification using multi-scale convolutional network. *BMC Bioinf.* 20, 730. <https://doi.org/10.1186/s12859-019-3327-y>.
22. Fu, H., Cao, Z., Li, M., and Wang, S. (2020). ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding. *BMC Genom.* 21, 597. <https://doi.org/10.1186/s12864-020-06978-0>.
23. Yi, H.-C., You, Z.-H., Zhou, X., Cheng, L., Li, X., Jiang, T.-H., and Chen, Z.H. (2019). ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. *Mol. Ther. Nucleic Acids* 17, 1–9. <https://doi.org/10.1016/j.omtn.2019.04.025>.
24. Ahmed, S., Muhammad, R., Khan, Z.H., Adilina, S., Sharma, A., Shatabda, S., and Dehzangi, A. (2021). ACP-MHCNN: an accurate multi-headed deep-convolutional neural network to predict anticancer peptides. *Sci. Rep.* 11, 23676. <https://doi.org/10.1038/s41598-021-02703-3>.
25. Wang, H., Zhao, J., Zhao, H., Li, H., and Wang, J. (2021). CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model. *BMC Bioinf.* 22, 512. <https://doi.org/10.1186/s12859-021-04433-9>.
26. Lv, Z., Cui, F., Zou, Q., Zhang, L., and Xu, L. (2021). Anticancer peptides prediction with deep representation learning features. *Briefings Bioinf.* 22, bbab008. <https://doi.org/10.1093/bib/bbab008>.
27. Galvão, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C., and Saldanha, T.C.B. (2005). A method for calibration and validation subset partitioning. *Talanta* 67, 736–740. <https://doi.org/10.1016/j.talanta.2005.03.025>.
28. Liu, W., Zhao, Z., Yuan, H.-F., Song, C.-F., and Li, X.Y. (2014). An optimal selection method of samples of calibration set and validation set for spectral multivariate analysis. *Spectrosc. Spectr. Anal.* 34, 947–951. [https://doi.org/10.3964/j.issn.1000-0593\(2014\)04-0947-05](https://doi.org/10.3964/j.issn.1000-0593(2014)04-0947-05).
29. Gao, T., Hu, L., Jia, Z., Xia, T., Fang, C., Li, H., Hu, L., Lu, Y., and Li, H. (2019). SPXYE: an improved method for partitioning training and validation sets. *Cluster Comput.* 22, 3069–3078. <https://doi.org/10.1007/s10586-018-1877-9>.
30. Lane, N., and Kahanda, I. (2020). DeepACPPred: A Novel Hybrid CNN-RNN Architecture for Predicting Anti-cancer Peptides (Springer), pp. 60–69. [https://doi.org/10.1007/978-3-030-54568-0\\_7](https://doi.org/10.1007/978-3-030-54568-0_7).
31. Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., and Raghava, G.P. (2021). AntiCP 2.0: an updated model for predicting anticancer peptides. *Briefings Bioinf.* 22, bbaa153. <https://doi.org/10.1093/bib/bbaa153>.
32. Thomas, S., Karnik, S., Barai, R.S., Jayaraman, V.K., and Ilicula-Thomas, S. (2010). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res.* 38, D774–D780. <https://doi.org/10.1093/nar/gkp1021>.
33. Harrington, P.B. (2018). Multiple versus single set validation of multivariate models to avoid mistakes. *Crit. Rev. Anal. Chem.* 48, 33–46. <https://doi.org/10.1080/10408347.2017.1361314>.
34. Chen, T., and Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System, pp. 785–794. <https://doi.org/10.1145/2939972.2939785>.
35. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation Networks, pp. 7132–7141. <https://doi.org/10.1109/TPAMI.2019.2913372>.
36. Woo, S., Park, J., Lee, J.-Y., and Kweon, I.S. (2018). Cbam: Convolutional Block Attention Module, pp. 3–19. [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1).
37. Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H., and Chou, K.C. (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* 436, 168–177. <https://doi.org/10.1016/j.ab.2013.01.019>.
38. Li, C., Sutherland, D., Hammond, S.A., Yang, C., Taho, F., Bergman, L., Houston, S., Warren, R.L., Wong, T., Hoang, L.M.N., et al. (2022). AMPLify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genom.* 23, 77. <https://doi.org/10.1186/s12864-022-08310-4>.
39. Mu, Z., Yu, T., Qi, E., Liu, J., and Li, G. (2019). DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information. *BMC Bioinf.* 20, 351. <https://doi.org/10.1186/s12859-019-2943-x>.
40. Jeffrey, H.J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res.* 18, 2163–2170. <https://doi.org/10.1093/nar/18.8.2163>.
41. Wan, Y., Wang, Z., and Lee, T.Y. (2021). Incorporating support vector machine with sequential minimal optimization to identify anticancer peptides. *BMC Bioinf.* 22, 286. <https://doi.org/10.1186/s12859-021-03965-4>.
42. Oda, T., Lim, K., and Tomii, K. (2017). Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinf.* 18, 288. <https://doi.org/10.1186/s12859-017-1686-9>.
43. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

- Nucleic Acids Res. 25, 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
44. Saha, I., Maulik, U., Bandyopadhyay, S., and Plewczynski, D. (2012). Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids* 43, 583–594. <https://doi.org/10.1007/s00726-011-1106-9>.
45. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
46. Lata, S., Sharma, B.K., and Raghava, G.P.S. (2007). Analysis and prediction of antibacterial peptides. *BMC Bioinf.* 8, 263. <https://doi.org/10.1186/1471-2105-8-263>.
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30. <https://doi.org/10.48550/arXiv.1706.03762>.

## Short article

# A deep learning platform to assess drug proarrhythmia risk

Ricardo Serrano,<sup>1,2,4</sup> Dries A.M. Feyen,<sup>1,2,4</sup> Arne A.N. Bruyneel,<sup>1,2,4</sup> Anna P. Hnatiuk,<sup>1,2</sup> Michelle M. Vu,<sup>1,2</sup> Prashila L. Amatya,<sup>1,2</sup> Isaac Perea-Gil,<sup>1,3</sup> Maricela Prado,<sup>1,3</sup> Timon Seeger,<sup>1,2</sup> Joseph C. Wu,<sup>1,2</sup> Ioannis Karakikes,<sup>1,3</sup> and Mark Mercola<sup>1,2,5,\*</sup>

<sup>1</sup>Stanford Cardiovascular Institute, Stanford University, Stanford, CA 94305, USA

<sup>2</sup>Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305, USA

<sup>3</sup>Department of Cardiothoracic Surgery, Stanford University, Stanford, CA 94305, USA

<sup>4</sup>These authors contributed equally

<sup>5</sup>Lead contact

\*Correspondence: mmercola@stanford.edu

<https://doi.org/10.1016/j.stem.2022.12.002>

## SUMMARY

Drug safety initiatives have endorsed human iPSC-derived cardiomyocytes (hiPSC-CMs) as an *in vitro* model for predicting drug-induced cardiac arrhythmia. However, the extent to which human-defined features of *in vitro* arrhythmia predict actual clinical risk has been much debated. Here, we trained a convolutional neural network classifier (CNN) to learn features of *in vitro* action potential recordings of hiPSC-CMs that are associated with lethal Torsade de Pointes arrhythmia. The CNN classifier accurately predicted the risk of drug-induced arrhythmia in people. The risk profile of the test drugs was similar across hiPSC-CMs derived from different healthy donors. In contrast, pathogenic mutations that cause arrhythmogenic cardiomyopathies in patients significantly increased the proarrhythmic propensity to certain intermediate and high-risk drugs in the hiPSC-CMs. Thus, deep learning can identify *in vitro* arrhythmic features that correlate with clinical arrhythmia and discern the influence of patient genetics on the risk of drug-induced arrhythmia.

## INTRODUCTION

Drug-induced arrhythmias are a common cause of drug attrition during development and for restricted use or withdrawal from the market.<sup>1–3</sup> As people vary in their predisposition to drug-induced arrhythmia,<sup>4–6</sup> there is a widely accepted need to assess risk in susceptible populations.<sup>7,8</sup> For ethical reasons and practical limitations, susceptible individuals, including carriers of rare predisposing gene variants, are not generally included in clinical trials.<sup>9</sup> Human iPSC-derived cardiomyocytes (hiPSC-CMs) retain an individual's genetic makeup and enable scalable production of cardiac cells for *in vitro* testing and, therefore, are a breakthrough technology for risk assessment.<sup>10,11</sup> This notion is supported by the findings that several mutations that cause electrophysiological or myopathic heart disease predispose CMs to drug-induced arrhythmia.<sup>12</sup>

Cell-based assays assess arrhythmia risk by quantifying waveform features in the cells' action potential. Typically, these features are quantified using human-defined metrics such as the action potential duration at 90% amplitude (APD<sub>90</sub>) or incidence of after depolarizations.<sup>6,13,14</sup> However, these human-defined features do not accurately predict clinical arrhythmia.<sup>15–17</sup> Altogether, the complex manifestations of arrhythmia, the uncertain correspondence between *in vitro* action potential waveform features and actual clinical arrhythmia, and the influence of dis-

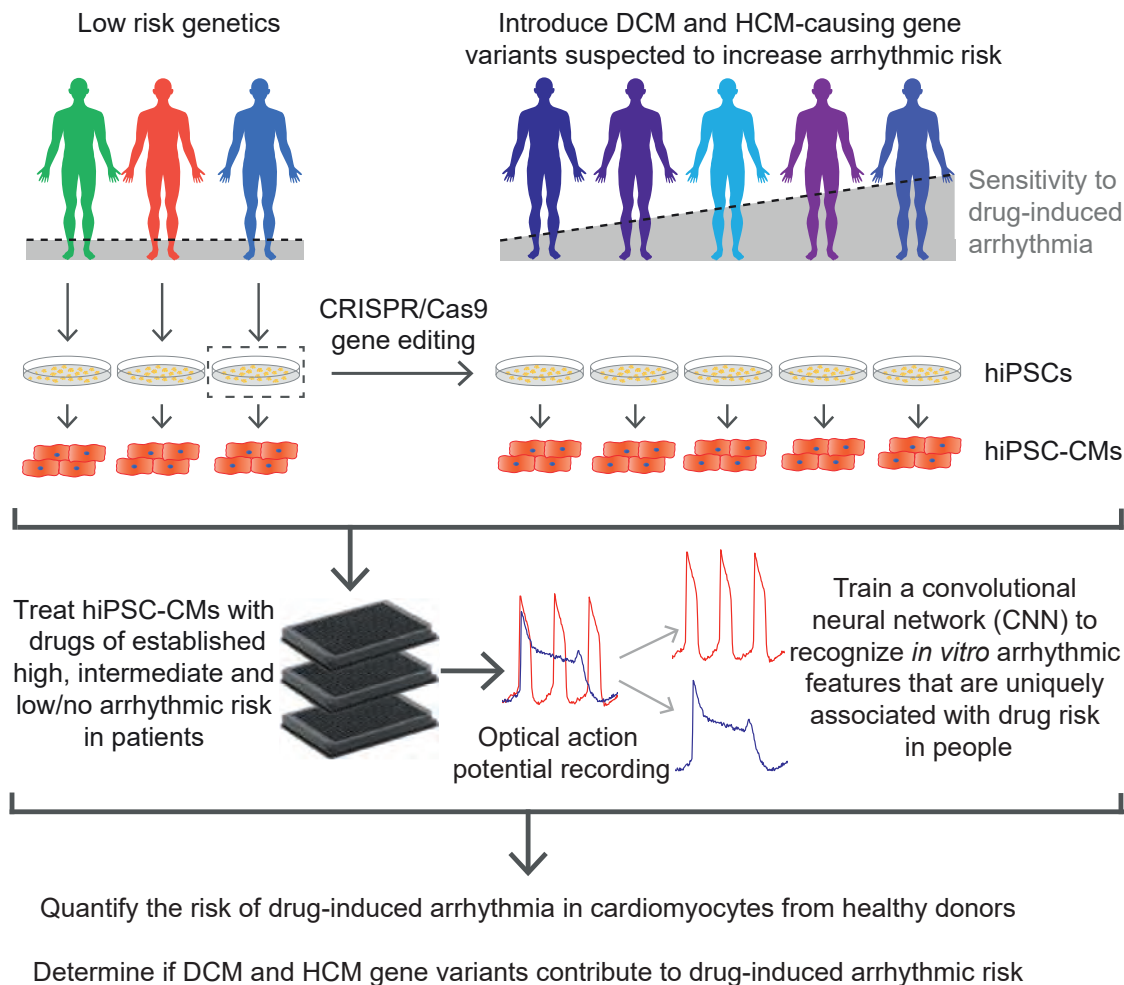
ease susceptibility loci present significant challenges for risk prediction.

To address this problem, we developed a deep learning approach to discriminate the *in vitro* electrophysiological features induced in hiPSC-CMs by reference drugs with well-established (high to low) risk of eliciting the life-threatening ventricular tachyarrhythmia (Torsades de Pointes, TdP). Deep learning is a type of artificial intelligence (AI) that uses multiple computational layers in a deep neural network (DNN). DNNs extract features relevant to discriminating input classes in a systematic and unbiased manner, effectively removing the need for human-defined metrics.<sup>18</sup> Among the different types of DNNs, convolutional neural networks (CNNs), which learn complex features from input data by assigning weights to the result of convolutional operations, are showing tremendous success in various biomedical applications such as medical image analysis<sup>19</sup> and physiological signal analysis.<sup>20</sup> Recently, CNNs have been used to automate the detection and classification of arrhythmias both *in vitro*<sup>21</sup> and *in vivo*.<sup>22</sup>

We trained a CNN to discriminate high versus low-risk drugs based on intrinsic drug-induced electrophysiological waveforms in hiPSC-CMs rather than human expectations. The CNN more accurately classified actual drug risk in patients than did human-defined metrics. Moreover, the trained CNN successfully quantified the increase risk of drug-induced arrhythmia caused







**Figure 1. Strategy to determine the influence of myopathic gene variants on the proarrhythmic effect of drugs**

hiPSC-CMs were generated from three donor patients without risk-associated genetics. DCM and HCM causing mutations were introduced to one of the healthy backgrounds and multiple batches of hiPSC-CMs were generated. The hiPSC-CMs were treated with 37 drugs of characterized high, intermediate, and low/no arrhythmic risk. Each drug was tested at 8 different concentrations and a voltage-sensitive dye was used to obtain membrane potential recordings. A CNN was trained to classify voltage traces based on the drug's risk of inducing arrhythmia in patients. Class probabilities from the CNN were used to rank the proarrhythmic effects of drugs and evaluate the influence of myopathic gene variants.

by cardiomyopathic gene variants, which pose a clinically significant risk factor that has been challenging to quantify.<sup>23</sup> In summary, deep learning out-performed human-defined methods for drug risk assessment and detected an influence of patient genetics on susceptibility to drug-induced arrhythmia.

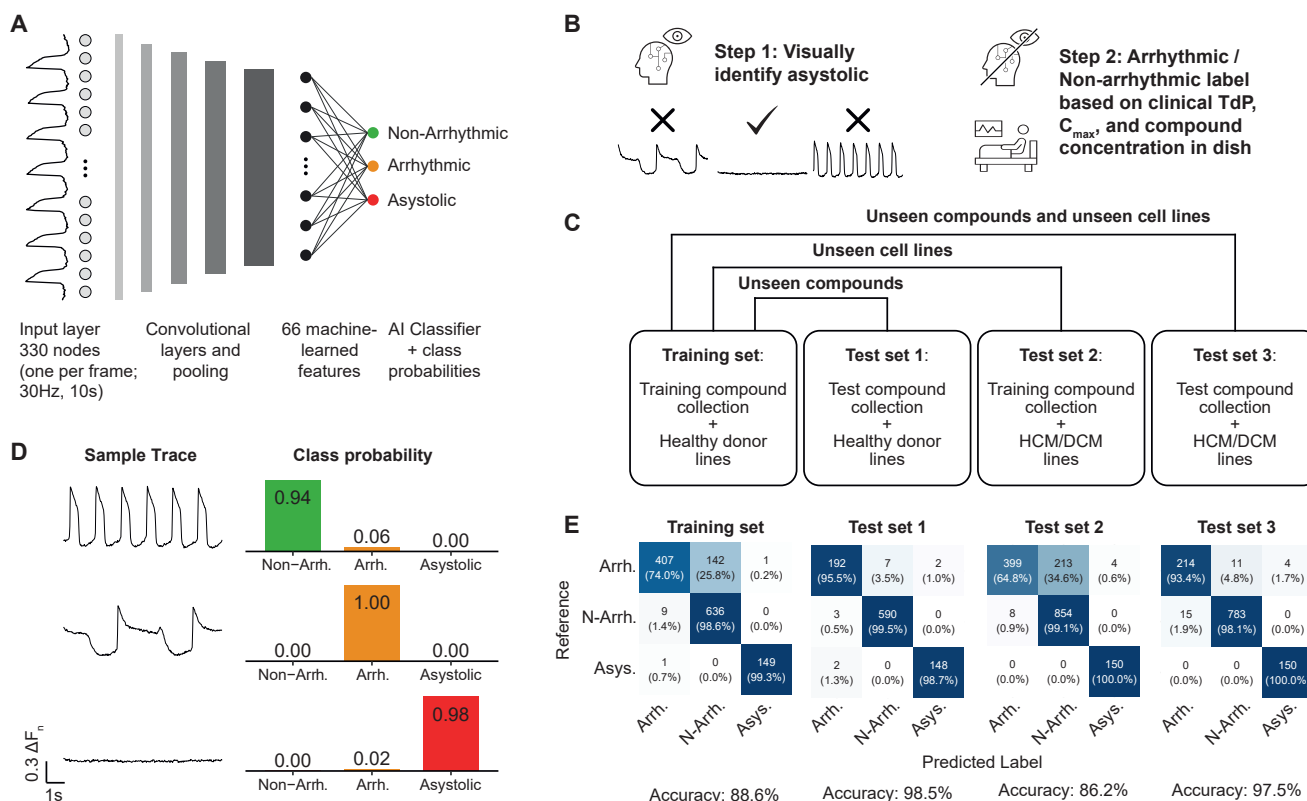
## RESULTS

### High-throughput screening of electrophysiological effects of drugs in healthy and disease hiPSC-CMs models

This study aimed to develop a new paradigm to accurately predict the risk of drugs and genetics on TdP arrhythmia without imposing human bias (Figure 1). Genotype-phenotype relationships were modeled in a cohort of hiPSC-CMs generated from 8 hiPSC lines. Three (3) lines were derived from healthy donors (HD.113, HD.273, and HD.15S1), while five (5) lines harbored

pathogenic mutations that cause hypertrophic cardiomyopathy [HCM] (HCM.MYBPC3 p.R943X),<sup>24</sup> left ventricular noncompaction (LVNC), and HCM (HCM.TPM1 p.K37E)<sup>25</sup> and dilated cardiomyopathy [DCM] (DCM.PLN p.R14del,<sup>26</sup> DCM.RBM20 p.R634Q,<sup>27</sup> and DCM.TNNT2 p.R183W<sup>28</sup>). These mutations were introduced in the same genetic background (HD.15S1 iPSCs) by CRISPR-mediated genome editing, minimizing the potentially confounding effects of the genetic background. All these mutations, except for the *TNNT2* p.R183W mutation, are associated with elevated arrhythmic risk in patients, including ventricular arrhythmias and sudden cardiac death.<sup>29–34</sup>

Experiments were carried out using a panel of drugs with well-characterized proarrhythmia risk profiles, including the 28 drugs proposed by comprehensive *in vitro* proarrhythmia assay (CiPA) initiative<sup>35</sup> and 9 additional drugs to facilitate generalization across studies. For training and evaluation purposes, we used the CiPA collection categorized as high, intermediate,



**Figure 2. A deep learning neural network for classification of voltage traces**

(A) Schematic of CNN classifier for voltage traces into the classes: non-arrhythmic, asystolic, and arrhythmic.  
 (B) Steps for trace annotation.  
 (C) Splitting data for training and test datasets.  
 (D) Examples of each trace category and values of class probabilities outputted by the CNN.  
 (E) Confusion matrices for the training and test datasets.

and low/no risk for developing TdP arrhythmia at clinical exposures based on published reports, the Food and Drug Administration (FDA) adverse effect database, and expert opinion.<sup>36</sup>

The uniformity and quality of the CM batches were assessed based on the baseline morphology of the action potential traces from 28 differentiation batches from 8 hiPSC lines displayed typical electrophysiological characteristics at baseline (see STAR Methods). Differentiation batches that presented with abnormal baseline characteristics, such as baseline arrhythmias and cessation in certain wells, were excluded from further analysis (Figure S2).

### Deep learning features from voltage traces

A CNN was designed to classify voltage traces as non-arrhythmic, arrhythmic, or asystolic (Figure 2A; see Figure S1A and STAR Methods for details). For training and testing the CNN, sets of annotated traces are required. First, traces with no spontaneous action potential activity were manually labeled as asystolic (Figure 2B). To remove human bias regarding the classification of non-arrhythmic versus arrhythmic traces, we employed the following criteria: (1) traces from wells treated with drugs carrying high risk of inducing TdP arrhythmia (classified according to CiPA)<sup>35</sup> and treated at a concentration greater or equal than maximum free plasma concentration (free  $C_{max}$ )

were annotated as arrhythmic. (2) Traces from wells treated with drugs with a no or low-risk CiPA classification and the concentration was less or equal to free  $C_{max}$  were annotated as non-arrhythmic. Traces that did not follow any of these criteria—i.e., wells treated with drugs of intermediate CiPA risk, wells treated with low doses of high-risk compounds, or high doses of low-risk compounds—were not included in either training or test datasets.

The training dataset comprised traces from healthy donor lines treated with the training compound library proposed by CiPA (Figures 2C and S1B). The training dataset was used to refine the convolutional layers' number and size of filters and dropout percentage until an accuracy of 88.6% was achieved. The confusion matrix of the training set (Figure 2E) revealed that most errors involved traces classified as “non-arrhythmic” by the CNN but were derived from drugs annotated as “arrhythmic.” Further analysis showed that 127 of these 142 misclassified traces were the result of treatment with bepridil, for which previous studies had shown consistently classified as producing non-arrhythmic responses in hiPSC-derived CMs despite being tested at concentrations exceeding a hundred times the free  $C_{max}$ .<sup>13,14,37</sup>

The accuracy of the trace classifier was verified against different sets of unseen data (Figures 2C and 2E). The test set 1 contained data from compounds that differed from the training

library but in the same cell lines used for training (yielding an accuracy of 98.5%). The test set 2 contained data from different cell lines than those used for training but the same compounds (yielding an accuracy of 86.2%). Lastly, test set 3 contained data from both cell lines and compounds that were not used for training (yielding an accuracy of 97.5%). Note that test set 2 contained unseen data from bepridil, which explains the lower accuracy relative to test sets 1 and 3.

Representative traces and the CNN class probabilities output are shown in Figure 2D. Briefly, the CNN interpreted whole voltage waveforms from non-arrhythmic, cessation, and arrhythmic classes and generated their respective probabilities (Figure 2D). These probabilities can be used as a unified set of metrics applicable for all action potential waveforms, overcoming the problems that human-defined metrics suffer when quantifying specific phenotypes (e.g., in asystole APD<sub>90</sub> loses its meaning as no action potential exists; or the detection of early after depolarizations [EADs] or delayed after depolarizations [DADs], which often requires human inspection of the trace). Most importantly, the training trace annotations were based on the risk of clinical arrhythmia. This approach allowed the CNN to learn features of the *in vitro* traces that corresponded to the effect of clinically risky drugs, circumventing the concern that human intuition might not recognize the *in vitro* features that are predictive of clinical arrhythmia.

### Use of the CNN to determine the probability of drug-induced arrhythmia

The trained CNN reflected the dose-dependent proarrhythmic effects in the voltage waveforms of CMs treated with high-risk drugs as increases in the probability of arrhythmic or asystolic phenotypes. To illustrate this result, Figure 3A shows traces from ibutilide treated hiPSC-CMs showing progressively arrhythmic traces with increasing dose (manifested by action potential prolongation and EADs). The trained CNN generated a continuous, dose-dependent increase in arrhythmia probability based on the waveform input (Figure 3B). Note that the calculated 50% arrhythmia class probability (EC<sub>50</sub>, triangle) was lower than the free C<sub>max</sub> value (dashed line), indicating that arrhythmia is detected at therapeutic concentrations of the drug. Consistent with the dependence of contractility on intracellular [Ca<sup>2+</sup>], the calcium channel blocker nifedipine showed a dose-dependent induction of asystole (Figure 3C) that corresponded with continuous increase in asystole probability determined by the CNN (Figure 3D). Similarly, dose-response curves can be calculated for any compound regardless of whether arrhythmic risk or even pharmacokinetic data are known.

### Quantitative metrics of drug safety

The preceding analysis was applied to the entire dataset (37 drugs, 8 hiPSC lines). Examples of CiPA-classified high (ibutilide), intermediate (ondansetron), and low (verapamil) arrhythmic risk compounds are shown with 3 drugs (flecainide, citalopram, and aspirin) that are not among the CiPA set but showed effects consistent with published clinical data (Figure 3E; Table S1; the entire dataset is shown in Data S1). In general, compounds classified by CiPA as having higher TdP risk displayed lower EC<sub>50</sub> values when normalized by C<sub>max</sub> (Figure 3F). To quantify this

observation, we elaborated the torsadogenic safety margin, defined as the log<sub>10</sub>(EC<sub>50</sub>/C<sub>max</sub>).

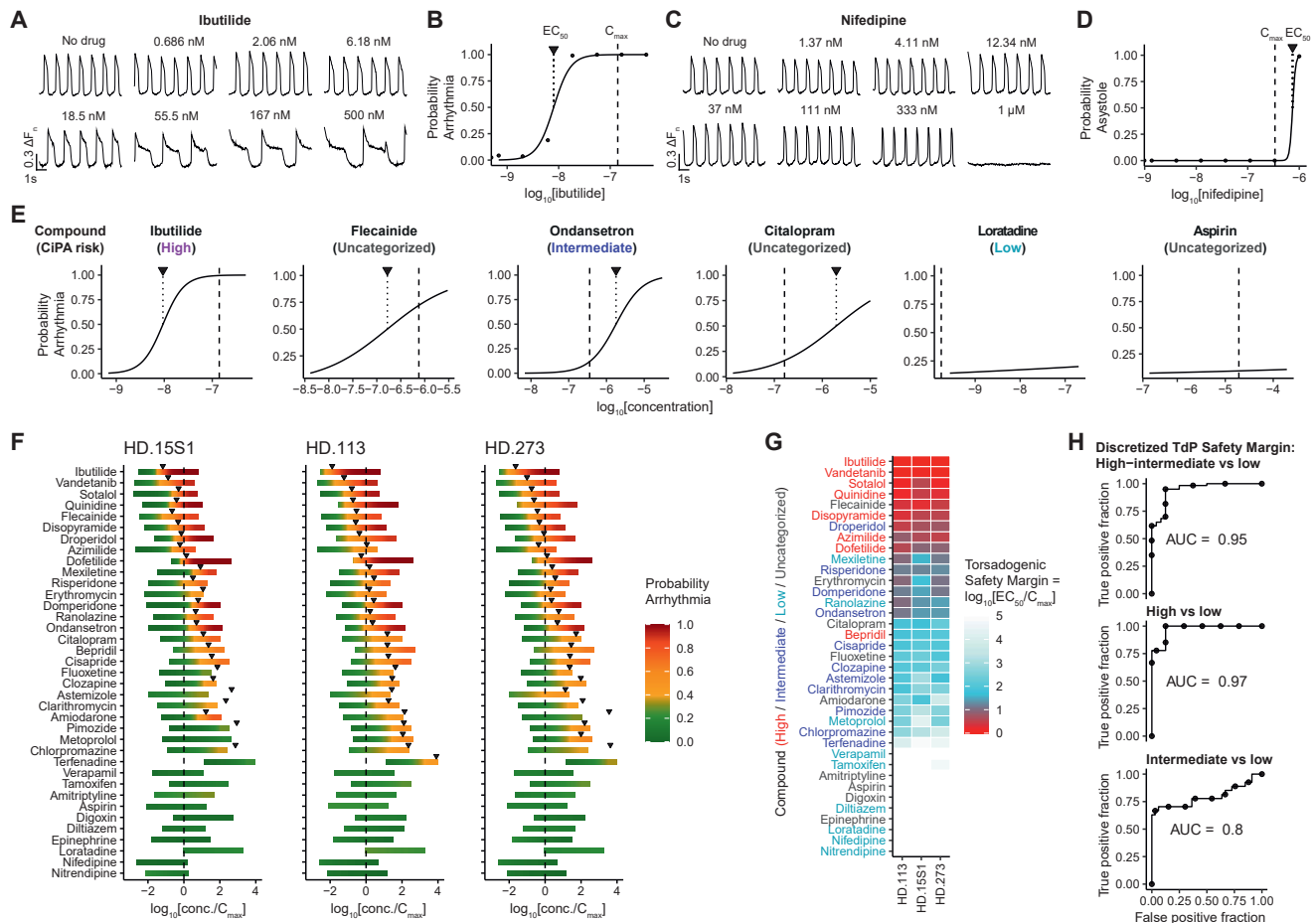
To visualize the relationship between the torsadogenic safety margin and clinical risk, we rank-ordered the compounds by their torsadogenic safety margin (Figure 3G). 8 out of the top 9 places were occupied by drugs classified by CiPA as carrying high risk (Table S1). The safety margin generally correlated with the CiPA-assigned clinical risk, although some exceptions were noted. Droperidol, classified as having intermediate risk by CiPA, ranked 7<sup>th</sup> highest, whereas bepridil, classified as high-risk by CiPA, was ranked 17<sup>th</sup>, placing it among intermediate-risk drugs. Mexiletine and ranolazine showed safety margin values comparable to intermediate-class drugs, as reported in other hiPSC-CMs studies,<sup>13,14,38</sup> despite being considered a low-no risk by CiPA classification.

The torsadogenic safety margin to classify risk can be used as a predictor in a logistic regression model to evaluate discretized CiPA risk. To compare to existing literature, we created a model to assign drugs to high-intermediate and low-no-risk discrete categories. This model resulted in an area under the curve (AUC) value of 0.95 (Figure 3H), higher than previously proposed models that use human-defined predictors applied to our dataset (Figure S3).<sup>13</sup> In addition, we trained models to distinguish high- and low-risk compounds (AUC = 0.97) and intermediate and low-risk compounds (AUC = 0.8) (Figure 3H).

We computed the torsadogenic safety margin of 9 additional drugs not contained in the CiPA collection (amiodarone, amitriptyline, aspirin, citalopram, digoxin, epinephrine, erythromycin, flecainide, and fluoxetine). We determined their risk by comparing their rank position against the CiPA reference compounds (Figure 3G). Flecainide ranked among high-risk drugs, whereas citalopram, fluoxetine, and amiodarone ranked as intermediate risk, and erythromycin was at the border between high and intermediate-risk. The probability of arrhythmia EC<sub>50</sub> values could not be determined for amitriptyline, aspirin, digoxin, and epinephrine, indicating that they are no- or low-risk drugs in agreement with the clinical literature on these drugs (Table S1). We conclude that the continuous nature of the torsadogenic safety margin allows it to stratify the risk of drugs at a finer scale than discrete categorization, in which all drugs within the same category (e.g., high risk) are considered equal.

### DCM and HCM mutations influence responses to proarrhythmic drugs

Next, we asked whether the CNN could discern the influence of patient genetics on drug-induced arrhythmia. To test this idea, we focused on variants that cause familial DCM and HCM. Certain DCM and HCM-causing variants place patients at risk for ventricular arrhythmias, and current treatment guidelines call for caution in treating familial DCM and HCM patients with torsadogenic drugs.<sup>39,40</sup> The mechanisms responsible for arrhythmia in familial DCM and HCM involve altered Ca<sup>2+</sup> and Na<sup>+</sup> flux as well as tissue remodeling (e.g., elevated fibrosis) in addition to K<sup>+</sup> current inhibition<sup>41–43</sup> that is responsible for most drug-induced TdP.<sup>44</sup> Despite these clinical and mechanistic associations, the contribution of DCM and HCM to drug-induced TdP risk has not been addressed quantitatively.



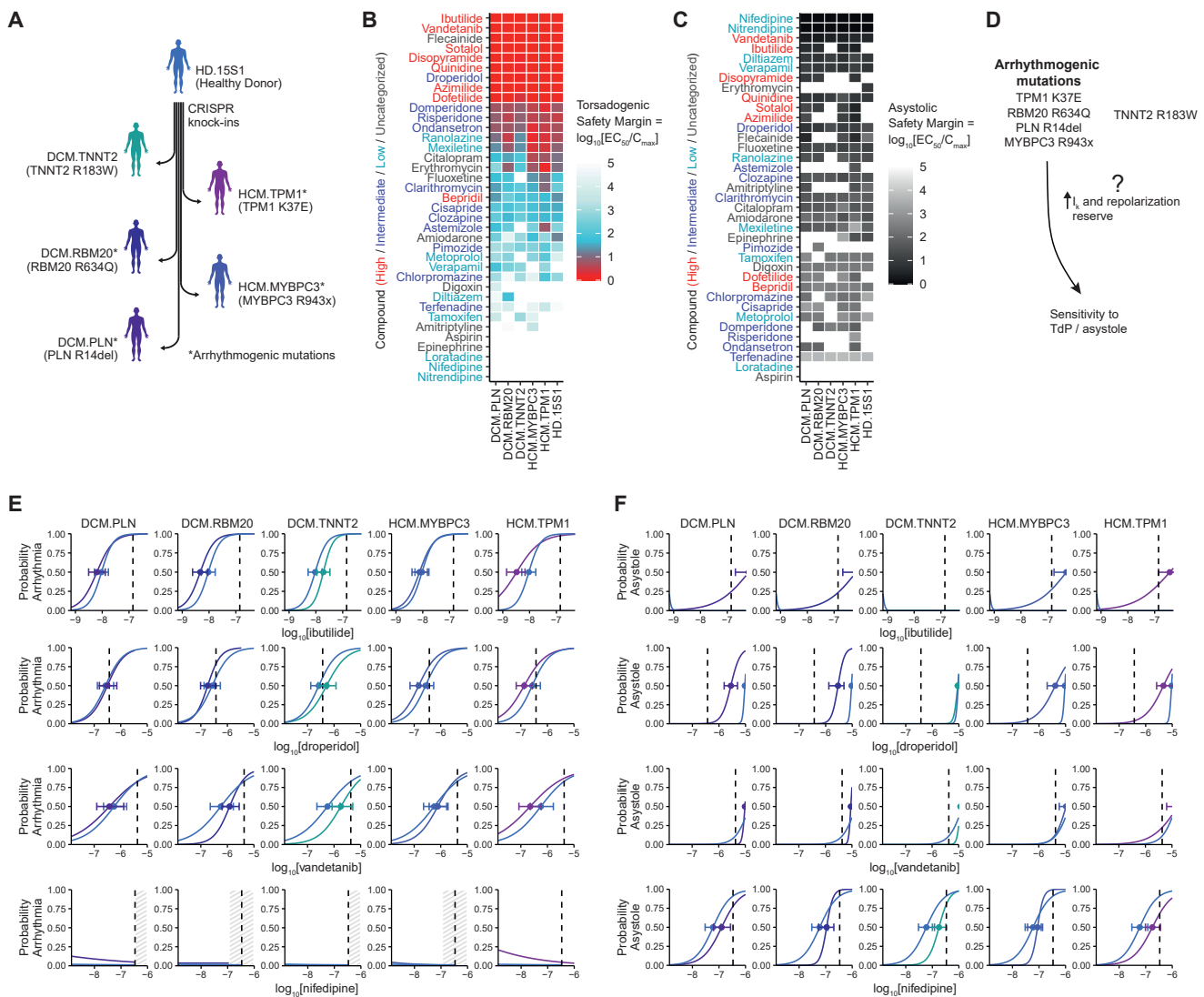
**Figure 3. Development of the torsadogenic safety margin**

(A) Representative traces of different concentrations of the torsadogenic drug ibutilide.  
 (B) Proarrhythmic dose-response curve (solid line) plotting the probability of arrhythmic class from the traces shown in (A). 50% probability of arrhythmic value  $EC_{50}$  (triangle). Clinical maximum free plasma concentration,  $C_{max}$  (dashed line).  
 (C) Representative traces of different concentrations of the calcium channel blocker nifedipine.  
 (D) Proarrhythmic dose-response curve (solid line) plotting the probability of asystole class from the traces shown in (A). 50% probability of asystole value  $EC_{50}$  (triangle). Clinical maximum free plasma concentration,  $C_{max}$  (dashed line). Average of 3 differentiation batches per cell line with 3 technical repeats per dose in each batch.  
 (E) Dose-response curves of for other drugs and their CiPA risk classification.  
 (F) Color-coded dose-response curves where probability of arrhythmic 0–1 is encoded as green-red, and the dose is normalized by  $C_{max}$ .  
 (G) Torsadogenic safety margin for each drug and healthy donor line of the screen. Drug names are color-coded based on CiPA classification.  
 (H) Receiver operating characteristic (ROC) curves for using the torsadogenic safety margin as predictor for a model to identify high-intermediate-risk drugs versus no-low, high versus no-low, and intermediate versus no-low.

Furthermore, an *in vitro* system to quantify TdP probability would aid in elucidating risk mechanisms.

We used the trained CNN to determine whether DCM and HCM mutations increase TdP probabilities in response to drug treatment. To remove the influence of background genetics, we created a panel of isogenic DCM and HCM lines by introducing disease-causing mutations into the healthy donor hiPSC line HD.15S1 (Figure 4A). The DCM-causing *RBM20* p.R634Q and *PLN* p.R14del variants and the HCM-causing *MYBPC3* p.R943X and *TPM1* p.K37E variants are recognized as “at risk” for fatal arrhythmia.<sup>29,30,32,33</sup> In contrast, the DCM *TNNT2* p.R183W is reported to confer less arrhythmic risk.<sup>34</sup>

Action potential waveforms in hiPSC-CMs derived from each line were recorded at baseline and upon dose-escalation treatment with the 37 drugs. Results in the mutant lines were compared with the isogenic control to determine the contribution of the gene variants. Most drugs showed similar torsadogenic safety margins in the DCM and HCM hiPSC-CMs as for the isogenic control, although some trended toward increased risk (e.g., ibutilide and droperidol) (Figure 4B and 4E). In contrast, the asystole safety margin (calculated analogously using the  $EC_{50}$  values of the probability of asystole) revealed a strong influence of DCM and HCM genotype on drug effects (Figure 4C). For example, ibutilide, dofetilide, and droperidol showed a



**Figure 4. Influence of DCM and HCM gene variants**

- (A) Myopathic gene variants were introduced by CRISPR-Cas9 gene editing onto a common healthy donor hiPSC line.
- (B) Torsadogenic safety margin for each drug and healthy donor line of the screen for the HCM and DCM cell lines and the healthy donor isogenic line. Drug names are color-coded based on CiPA classification.
- (C) Asystolic safety margin for each drug and healthy donor line of the screen for the HCM and DCM cell lines and the healthy donor isogenic line. Drug names are color-coded based on CiPA classification.
- (D) Hypothesized model for increased sensitivity of arrhythmogenic cell lines to TdP and asystole (see discussion).
- (E) Dose-response curves of probability of arrhythmic in HCM and DCM lines and the isogenic control cell lines, treated with ibutilide, droperidol, vandetanib, and nifedipine. Point markers indicate  $EC_{50}$  with error bars at a 95% confidence interval. Shaded region signifies all traces were asystolic at that concentration range. Average of 3 differentiation batches per cell line with 3 technical repeats per dose in each batch.
- (F) Dose-response curves of probability of asystolic in HCM and DCM lines and the isogenic control cell lines, treated with ibutilide, droperidol, vandetanib, and nifedipine. Point markers indicate  $EC_{50}$  with error bars at a 95% confidence interval. Average of 3 differentiation batches per cell line with 3 technical repeats per dose in each batch.

heightened propensity to cause asystole in the DCM variants *RBM20* p.R634Q and *PLN* p.R14del, and the HCM causal variants *MYBPC3* p.R943X, and *TPM1* p.K37E relative to the isogenic control hiPSC-CMs (Figures 4C and 4F). Interestingly, hiPSC-CMs carrying the DCM *TNNT2* p.R183W variant (that is less arrhythmogenic in patients) had a similar profile to the isogenic healthy donor control (Figures 4C and 4F).

## DISCUSSION

Current *in vitro* proarrhythmia assays rely on the measurement of human-defined features of cardiac electrophysiology, such as APD and beat rate.<sup>6,13,45</sup> However, arrhythmias present complex geometries, such as EADs, that are challenging to quantify by conventional metrics. In practice, arrhythmic phenotypes are

typically classified based on categorical descriptors (e.g., sustained ventricular tachycardia or after depolarizations) or binary variables (e.g., presence or absence of EADs) that require human interpretation of the action potential.<sup>13,14,38</sup> Treating arrhythmic phenotypes as yes/no parameters imposes an artificial threshold for an arrhythmia that depends on human intuition and cannot quantify progression from normal to arrhythmic waveforms as a function of drug dose, chemical modification, or hiPSC genetics. More fundamentally, human intuition is not based on an established ground truth regarding the *in vitro* effect of drugs that cause arrhythmia in patients. Thus, the motivation for this study was that human-defined metrics might not capture features of the *in vitro* waveforms that correspond with the actual arrhythmic risk of the drugs in people.

We developed our deep learning algorithm to overcome the limitations of human-defined metrics by recognizing features uniquely associated with drug-induced arrhythmia in hiPSC-CMs. Training a CNN to discriminate features was based on a dataset of 3 classes of action potential traces: (1) traces generated by treating hiPSC-CMs with high doses of drugs that cause ventricular arrhythmia in people (class: arrhythmic), (2) traces generated with low doses of safe compounds (class: non-arrhythmic), and (3) traces in which drugs induced asystole, which is characteristic of high doses of  $\text{Ca}^{2+}$  channel blockers (class: asystole) but also resulted from treatment with very high doses of proarrhythmic drugs (Figure 2). The probability of classification as arrhythmic and asystolic was a continuous dose-dependent metric that quantified the behavior of a drug. Relating the  $\text{EC}_{50}$  values for these probabilities to the free plasma concentration of a drug used in clinical practice (free  $C_{\text{max}}$ ) defined a torsadogenic safety margin for each drug (Figure 3). In hiPSC-CMs from healthy donors, the torsadogenic safety margin accurately predicted the clinical risk of drugs with an AUC of 0.95 (Figure 3H), representing an improvement over previously published multiparametric methods based on human classification (Figure S3).

Patients with structural heart diseases such as DCM and HCM are at risk for drug-induced arrhythmia. They should be carefully monitored using electrocardiographic and other modalities when treated with drugs at risk for inducing arrhythmia.<sup>40</sup> We applied the torsadogenic and asystolic safety margins to traces generated by treating isogenic hiPSC-CMs carrying gene variants that cause DCM and HCM in patients. The CNN probabilities and the calculated safety margins revealed that pathological gene variants associated with arrhythmic cardiomyopathies in patients sensitized hiPSC-CMs to adverse proarrhythmic and asystolic effects of high-risk drugs. In particular, hiPSC-CMs carrying pathogenic variants in *PLN*, *RBM20*, *TPM1*, and *MYBPC3* associated with increased arrhythmic risk in patients<sup>29,30,32,33</sup> trended toward increased probabilities of being classified as arrhythmic (Figures 4B and 4D) and showed highly significant increases in the probabilities of asystole (Figures 4C and 4E) compared with isogenic controls.

The finding that arrhythmogenic DCM and HCM gene variants increase the risk of torsadogenic drugs has implications for understanding the electrophysiological substrates for arrhythmia. The genetic lesions examined here affect sarcomeric, RNA splicing, and  $\text{Ca}^{2+}$  handling proteins. Electrophysiological remodeling in familial DCM and HCM includes reduced repolarizing

$\text{K}^+$  currents ( $I_{\text{K}}$ ) and increased intracellular diastolic  $\text{Ca}^{2+}$  and late  $\text{Na}^+$  ( $I_{\text{NaL}}$ ) current, as reviewed.<sup>42,43</sup> For example, these current changes have been reported in isolated CMs from mice, human myectomy samples, and hiPSCs carrying *MYBPC3* mutations that are functionally equivalent to the p.R943X truncation mutation used in this study.<sup>41,46–48</sup> A downstream circuit involving calmodulin-dependent protein kinase II CaMKII sustains electrophysiological remodeling<sup>49,50</sup> and decreases repolarizing  $I_{\text{K}}$ .<sup>51</sup> Decreased  $I_{\text{K}}$  in the hiPSC-CMs, which are more depolarized relative to adult CMs (at least in monolayer culture),<sup>52,53</sup> would enhance the effect of torsadogenic drugs<sup>44</sup> and is a possible mechanism for their asystolic effect in hiPSC-CMs carrying the DCM and HCM mutations (Figure 4D).

In conclusion, the deep learning algorithm recognized *in vitro* arrhythmic features in the hiPSC-CMs based on drug effects. It did not rely on human adjudication nor human-defined *in vitro* hiPSC-CM arrhythmic phenotypes; instead, it focused the CNN on recognizing hiPSC-CM phenotypes associated with patient drug responses. We derived safety margins from the relationship between the machine-generated class probabilities and the free plasma concentrations of each drug. Unlike categorical descriptors of arrhythmia (such as EADs, action potential prolongation, and triangulation), the CNN class probabilities map each trace to a continuous spectrum spanning non-arrhythmic to arrhythmic and asystolic phenotypes. Overall, the calculated safety margins more accurately discriminated high-, intermediate-, and low-risk drugs than prior methods based on human-defined features and revealed arrhythmogenic HCM and DCM variants increase sensitivity to drug-induced arrhythmia. Thus, the recognition of hiPSC-CM features of arrhythmia by deep learning should improve the detection of risky compounds during development as well as assign risk to gene variants of unknown significance.

### Limitations of the study

Deep learning has the inherent limitation that the machine-learned features typically lack human-intelligible meaning (the so-called “black box” problem). In other words, the mathematical operations that the machine has optimized to calculate the probability of arrhythmia are too abstract to be interpretable by humans and cannot be ascribed to action potential features. Although we cannot query the CNN to understand the basis for classification, visual inspection of the traces revealed that the dose-dependent increases in arrhythmia probability corresponded to action potential (AP) prolongation, EADs, and alternans. The CiPA reference collection causes a limited range of arrhythmogenic phenotypes (typically AP prolongation and EADs). We have found that the CNN (R.S., unpublished data) can be applied to a broader range of phenotypes (e.g., DADs, sustained and non-sustained ventricular tachycardia associated with cardiac electrophysiological disorders)<sup>54–56</sup> by increasing the number of possible classifications (i.e., the number nodes in the last layer of the CNN).

Certain drugs in this study (droperidol, ranolazine, and mexiletine) presented as “riskier” than in actual clinical practice while bepridil appeared less “risky” as discussed above. Previous studies using hiPSC-CMs similarly misclassified these drugs.<sup>13,37</sup> This might reflect a limitation in using hiPSC-CM datasets to classify these drugs. Concerning the deep learning

methodology, we speculate that the features needed to accurately classify these drugs might have been masked by variability in the waveforms in the current dataset and that a larger dataset might be adequately powered to train the neural network to correctly define the risk of these drugs.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - hiPSCs culture and differentiation
  - Genome editing in hiPSCs
- METHOD DETAILS
  - High throughput optical voltage assay
  - Convolutional neural network classifier
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Probability of arrhythmic and asystolic dose-response curves
  - Calculation of AUC of discretized TdP safety margin

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.stem.2022.12.002>.

## ACKNOWLEDGMENTS

This research was made possible by grants from the National Institutes of Health (R01HL130840, R01HL138539, R01HL141358, 1R01HL152055, 1R42HL158510, and Fondation Leducq 18CVD01 to M.M., P01HL141084 to M.M. and J.C.W., American Heart Association 17MERIT33610009 and Fondation Leducq 18CVD05 to J.C.W. and R01HL150414 and R01HL139679 to I.K.), and the Joan and Sanford I. Weill Scholars Endowment. D.A.M.F. was funded by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 708459. Graphical abstract was created with BioRender.com.

## AUTHOR CONTRIBUTIONS

Conceptualization, R.S., D.A.M.F., A.N.N.B., I.K., and M.M.; methodology, R.S., D.A.M.F., A.N.N.B., A.P.H., I.P.-G., M.M.V., P.L.A., and M.P.; investigation, R.S., D.A.M.F., A.N.N.B., I.P.-G., M.M.V., P.L.A., M.P., and T.S.; writing, R.S., D.A.M.F., A.N.N.B., A.P.H., and M.M.; review & editing, all authors; funding acquisition, I.K., J.C.W., and M.M.; supervision, J.C.W., I.K., and M.M.

## DECLARATION OF INTERESTS

R.S. is a paid consultant of Vala Sciences, which manufactures a high content instrument used in these studies. M.M. serves on the scientific advisory board of Vala Sciences. J.C.W. is co-founder and scientific advisory board member of Greenstone Biosciences.

Received: June 24, 2022

Revised: October 25, 2022

Accepted: November 29, 2022

Published: December 22, 2022

## REFERENCES

1. Valentin, J.-P., and Delaunois, A. (2018). Developing solutions to detect and avoid cardiovascular toxicity in the clinic. *Toxicology Letters* 295, S48–S49.
2. Onakpoya, I.J., Heneghan, C.J., and Aronson, J.K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med.* 14, 10. <https://doi.org/10.1186/s12916-016-0553-2>.
3. Waring, M.J., Arrowsmith, J., Leach, A.R., Leeson, P.D., Mandrell, S., Owen, R.M., Pairedeau, G., Pennie, W.D., Pickett, S.D., Wang, J., et al. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* 14, 475–486. <https://doi.org/10.1038/nrd4609>.
4. Baguley, W.A., Hay, W.T., Mackie, K.P., Cheney, F.W., and Cullen, B.F. (1997). Cardiac dysrhythmias associated with the intravenous administration of ondansetron and metoclopramide. *Anesth. Analg.* 84, 1380–1381. <https://doi.org/10.1097/0000539-199706000-00038>.
5. Frommeyer, G., and Eckardt, L. (2016). Drug-induced proarrhythmia: risk factors and electrophysiological mechanisms. *Nat. Rev. Cardiol.* 13, 36–47. <https://doi.org/10.1038/nrcardio.2015.110>.
6. Liang, P., Lan, F., Lee, A.S., Gong, T., Sanchez-Freire, V., Wang, Y., Diecke, S., Sallam, K., Knowles, J.W., Wang, P.J., et al. (2013). Drug screening using a library of human induced pluripotent stem cell-derived cardiomyocytes reveals disease-specific patterns of cardiotoxicity. *Circulation* 127, 1677–1691. <https://doi.org/10.1161/CIRCULATIONAHA.113.001883>.
7. Laverty, H., Benson, C., Cartwright, E., Cross, M., Garland, C., Hammond, T., Holloway, C., McMahon, N., Milligan, J., Park, B., et al. (2011). How can we improve our understanding of cardiovascular safety liabilities to develop safer medicines? *Br. J. Pharmacol.* 163, 675–693. <https://doi.org/10.1111/j.1476-5381.2011.01255.x>.
8. Gintant, G., Sager, P.T., and Stockbridge, N. (2016). Evolution of strategies to improve preclinical cardiac safety testing. *Nat. Rev. Drug Discov.* 15, 457–471. <https://doi.org/10.1038/nrd.2015.34>.
9. Heist, E.K., and Ruskin, J.N. (2010). Drug-induced arrhythmia. *Circulation* 122, 1426–1435. <https://doi.org/10.1161/CIRCULATIONAHA.109.894725>.
10. Savoji, H., Mohammadi, M.H., Rafatian, N., Toroghi, M.K., Wang, E.Y., Zhao, Y., Korolj, A., Ahadian, S., and Radisic, M. (2019). Cardiovascular disease models: a game changing paradigm in drug discovery and screening. *Biomaterials* 198, 3–26. <https://doi.org/10.1016/j.biomaterials.2018.09.036>.
11. Paik, D.T., Chandy, M., and Wu, J.C. (2020). Patient and disease-specific induced pluripotent stem cells for discovery of personalized cardiovascular drugs and therapeutics. *Pharmacol. Rev.* 72, 320–342. <https://doi.org/10.1124/pr.116.013003>.
12. Hnatiuk, A.P., Briganti, F., Staudt, D.W., and Mercola, M. (2021). Human iPSC modeling of heart disease for drug development. *Cell Chem. Biol.* 28, 271–282.
13. Blinova, K., Dang, Q., Millard, D., Smith, G., Pierson, J., Guo, L., Brock, M., Lu, H.R., Kraushaar, U., Zeng, H., et al. (2018). International multisite study of human-induced pluripotent stem cell-derived cardiomyocytes for drug proarrhythmic potential assessment. *Cell Rep.* 24, 3582–3592. <https://doi.org/10.1016/j.celrep.2018.08.079>.
14. Pfeiffer, E.R., Vega, R., McDonough, P.M., Price, J.H., and Whittaker, R. (2016). Specific prediction of clinical QT prolongation by kinetic image cytometry in human stem cell derived cardiomyocytes. *J. Pharmacol. Toxicol. Methods* 81, 263–273. <https://doi.org/10.1016/j.vascn.2016.04.007>.
15. Antzelevitch, C. (2004). Arrhythmogenic mechanisms of QT prolonging drugs: is QT prolongation really the problem? *J. Electrocardiol.* 37 (Suppl.), 15–24.
16. Vicente, J., Stockbridge, N., and Strauss, D.G. (2016). Evolving regulatory paradigm for proarrhythmic risk assessment for new drugs.

- J. *Electrocardiol.* 49, 837–842. <https://doi.org/10.1016/j.jelectrocard.2016.07.017>.
17. Bhuiyan, T.A., Graff, C., Kanters, J.K., Melgaard, J., Toft, E., Kääh, S., and Struijk, J.J. (2018). A history of drug-induced torsades de pointes is associated with T-wave morphological abnormalities. *Clin. Pharmacol. Ther.* 103, 1100–1106. <https://doi.org/10.1002/cpt.886>.
  18. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. <https://doi.org/10.1038/nature14539>.
  19. Ker, J., Wang, L., Rao, J., and Lim, T. (2018). Deep learning applications in medical image analysis. *IEEE Access* 6, 9375–9389. <https://doi.org/10.1109/ACCESS.2017.2788044>.
  20. Faust, O., Hagiwara, Y., Hong, T.J., Lih, O.S., and Acharya, U.R. (2018). Deep learning for healthcare applications based on physiological signals: a review. *Comput. Methods Programs Biomed.* 167, 1–13. <https://doi.org/10.1016/j.cmpb.2018.04.005>.
  21. Golgooni, Z., Mirsadeghi, S., Soleymani Baghshah, M., Ataee, P., Baharvand, H., Pahlavan, S., and Rabiee, H.R. (2019). Deep learning-based proarrhythmia analysis using field potentials recorded from human pluripotent stem cells derived cardiomyocytes. *IEEE J. Transl. Eng. Health Med.* 7, 1–9. <https://doi.org/10.1109/JTEHM.2019.2907945>.
  22. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M., Tison, G.H., Bourn, C., Turakhia, M.P., and Ng, A.Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* 25, 65–69. <https://doi.org/10.1038/s41591-018-0268-3>.
  23. Petropoulou, E., Jamshidi, Y., and Behr, E.R. (2014). The genetics of proarrhythmic adverse drug reactions. *Br. J. Clin. Pharmacol.* 77, 618–625. <https://doi.org/10.1111/bcp.12208>.
  24. Seeger, T., Shrestha, R., Lam, C.K., Chen, C., McKeithan, W.L., Lau, E., Wnorowski, A., McMullen, G., Greenhaw, M., Lee, J., et al. (2019). A premature termination codon mutation in MYBPC3 causes hypertrophic cardiomyopathy via chronic activation of nonsense-mediated decay. *Circulation* 139, 799–811. <https://doi.org/10.1161/CIRCULATIONAHA.118.034624>.
  25. Perea Gil, I., Bellbachir, N., Gavidia, A.A., Arthur, J., Zhang, Y., Vadgama, N., Oikonomopoulos, A., Roura, S., Wu, J.C., Bayes-Genis, A., and Karakikes, I. (2020). Abstract 274: activation of CaMKII signaling pathway contributes to the pathogenesis of genetic hypertrophic cardiomyopathy. *Circ. Res.* 127, A274. [https://doi.org/10.1161/res.127.suppl\\_1.274](https://doi.org/10.1161/res.127.suppl_1.274).
  26. Feyen, D.A.M., Perea-Gil, I., Maas, R.G.C., Harakalova, M., Gavidia, A.A., Arthur Ataam, J., Wu, T.H., Vink, A., Pei, J., Vadgama, N., et al. (2021). Unfolded protein response as a compensatory mechanism and potential therapeutic target in PLN R14del cardiomyopathy. *Circulation* 144, 382–392. <https://doi.org/10.1161/CIRCULATIONAHA.120.049844>.
  27. Briganti, F., Sun, H., Wei, W., Wu, J., Zhu, C., Liss, M., Karakikes, I., Rego, S., Cipriano, A., Snyder, M., et al. (2020). iPSC modeling of RBM20-deficient DCM identifies upregulation of RBM20 as a therapeutic strategy. *Cell Rep.* 32, 108117. <https://doi.org/10.1016/j.celrep.2020.108117>.
  28. Perea-Gil, I., Seeger, T., Bruyneel, A.A.N., Termglinchan, V., Monte, E., Lim, E.W., Vadgama, N., Furihata, T., Gavidia, A.A., Arthur Ataam, J., et al. (2022). Serine biosynthesis as a novel therapeutic target for dilated cardiomyopathy. *Eur. Heart J.* 43, 3477–3489. <https://doi.org/10.1093/eurheartj/ehac305>.
  29. Niimura, H., Bachinski, L.L., Sangwatanaroj, S., Watkins, H., Chudley, A.E., McKenna, W., Kristinsson, A., Roberts, R., Sole, M., Maron, B.J., et al. (1998). Mutations in the gene for cardiac myosin-binding protein C and late-onset familial hypertrophic cardiomyopathy. *N. Engl. J. Med.* 338, 1248–1257.
  30. Chang, B., Nishizawa, T., Furutani, M., Fujiki, A., Tani, M., Kawaguchi, M., Ibuki, K., Hirano, K., Taneichi, H., Uese, K., et al. (2011). Identification of a novel TPM1 mutation in a family with left ventricular noncompaction and sudden death. *Mol. Genet. Metab.* 102, 200–206. <https://doi.org/10.1016/j.ymgme.2010.09.009>.
  31. Haghighi, K., Kolokathis, F., Gramolini, A.O., Waggoner, J.R., Pater, L., Lynch, R.A., Fan, G.C., Tsiapras, D., Parekh, R.R., Dorn, G.W., 2nd, et al. (2006). A mutation in the human phospholamban gene, deleting arginine 14, results in lethal, hereditary cardiomyopathy. *Proc. Natl. Acad. Sci. USA.* 103, 1388–1393. <https://doi.org/10.1073/pnas.0510519103>.
  32. van der Zwaag, P.A., van Rijsingen, I.A.W., Asimaki, A., Jongbloed, J.D.H., van Veldhuisen, D.J., Wiesfeld, A.C.P., Cox, M.G.P.J., van Lochem, L.T., de Boer, R.A., Hofstra, R.M.W., et al. (2012). Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: evidence supporting the concept of arrhythmogenic cardiomyopathy. *Eur. J. Heart Fail.* 14, 1199–1207. <https://doi.org/10.1093/eurjhf/hfs119>.
  33. Parikh, V.N., Caleshu, C., Reuter, C., Lazzeroni, L.C., Ingles, J., Garcia, J., McCaleb, K., Adesiyun, T., Sedaghat-Hamedani, F., Kumar, S., et al. (2019). Regional variation in RBM20 causes a highly penetrant arrhythmogenic cardiomyopathy. *Circ. Heart Fail.* 12, e005371. <https://doi.org/10.1161/CIRCHEARTFAILURE.118.005371>.
  34. Campbell, N., Sinagra, G., Jones, K.L., Slavov, D., Gowan, K., Merlo, M., Carniel, E., Fain, P.R., Aragona, P., Di Lenarda, A., et al. (2013). Whole exome sequencing identifies a troponin T mutation hot spot in familial dilated cardiomyopathy. *PLoS One* 8, e78104. <https://doi.org/10.1371/journal.pone.0078104>.
  35. Sager, P.T., Gintant, G., Turner, J.R., Pettit, S., and Stockbridge, N. (2014). Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium. *Am. Heart J.* 167, 292–300. <https://doi.org/10.1016/j.ahj.2013.11.004>.
  36. Colatsky, T., Fermini, B., Gintant, G., Pierson, J.B., Sager, P., Sekino, Y., Strauss, D.G., and Stockbridge, N. (2016). The comprehensive in vitro proarrhythmia assay (CiPA) initiative - update on progress. *J. Pharmacol. Toxicol. Methods* 87, 15–20. <https://doi.org/10.1016/j.vascn.2016.06.002>.
  37. Ando, H., Yoshinaga, T., Yamamoto, W., Asakura, K., Uda, T., Taniguchi, T., Ojima, A., Shinkyo, R., Kikuchi, K., Osada, T., et al. (2017). A new paradigm for drug-induced torsadogenic risk assessment using human iPSC cell-derived cardiomyocytes. *J. Pharmacol. Toxicol. Methods* 84, 111–127. <https://doi.org/10.1016/j.vascn.2016.12.003>.
  38. Kitaguchi, T., Moriyama, Y., Taniguchi, T., Ojima, A., Ando, H., Uda, T., Otabe, K., Oguchi, M., Shimizu, S., Saito, H., et al. (2016). CSAHI study: evaluation of multi-electrode array in combination with human iPSC cell-derived cardiomyocytes to predict drug-induced QT prolongation and arrhythmia-effects of 7 reference compounds at 10 facilities. *J. Pharmacol. Toxicol. Methods* 78, 93–102. <https://doi.org/10.1016/j.vascn.2015.12.002>.
  39. Maron, B.J., Rowin, E.J., and Maron, M.S. (2022). Hypertrophic cardiomyopathy: new concepts and therapies. *Annu. Rev. Med.* 73, 363–375. <https://doi.org/10.1146/annurev-med-042220-021539>.
  40. Tisdale, J.E., Chung, M.K., Campbell, K.B., Hammadah, M., Joglar, J.A., Leclerc, J., and Rajagopalan, B.; American Heart Association Clinical Pharmacology Committee of the Council on Clinical Cardiology and Council on Cardiovascular and Stroke Nursing (2020). Drug-induced arrhythmias: a scientific statement from the American Heart Association. *Circulation* 142, e214–e233. <https://doi.org/10.1161/CIR.0000000000000905>.
  41. Toib, A., Zhang, C., Borghetti, G., Zhang, X., Wallner, M., Yang, Y., Troupes, C.D., Kubo, H., Sharp, T.E., Feldsott, E., et al. (2017). Remodeling of repolarization and arrhythmia susceptibility in a myosin-binding protein C knockout mouse model. *Am. J. Physiol. Heart Circ. Physiol.* 313, H620–H630. <https://doi.org/10.1152/ajpheart.00167.2017>.
  42. Santini, L., Coppini, R., and Cerbai, E. (2021). Ion channel impairment and myofilament Ca<sup>2+</sup> sensitization: two parallel mechanisms underlying arrhythmogenesis in hypertrophic cardiomyopathy. *Cells* 10, 2789. <https://doi.org/10.3390/cells10102789>.
  43. Maron, B.J., Rowin, E.J., and Maron, M.S. (2019). Paradigm of sudden death prevention in hypertrophic cardiomyopathy. *Circ. Res.* 125, 370–378. <https://doi.org/10.1161/CIRCRESAHA.119.315159>.
  44. Roden, D.M. (2008). Cellular basis of drug-induced torsades de pointes. *Br. J. Pharmacol.* 154, 1502–1507. <https://doi.org/10.1038/bjp.2008.238>.



45. McKeithan, W.L., Savchenko, A., Yu, M.S., Cerignoli, F., Bruyneel, A.A.N., Price, J.H., Colas, A.R., Miller, E.W., Cashman, J.R., and Mercola, M. (2017). An automated platform for assessment of congenital and drug-induced arrhythmia with hiPSC-derived cardiomyocytes. *Front. Physiol.* **8**, 766. <https://doi.org/10.3389/fphys.2017.00766>.
46. Coppini, R., Ferrantini, C., Yao, L., Fan, P., Del Lungo, M., Stillitano, F., Sartiani, L., Tosi, B., Suffredini, S., Tesi, C., et al. (2013). Late sodium current inhibition reverses electromechanical dysfunction in human hypertrophic cardiomyopathy. *Circulation* **127**, 575–584. <https://doi.org/10.1161/CIRCULATIONAHA.112.134932>.
47. Flenner, F., Jungen, C., Küpker, N., Ibel, A., Kruse, M., Koivumäki, J.T., Rinas, A., Zech, A.T.L., Rhoden, A., Wijmker, P.J.M., et al. (2021). Translational investigation of electrophysiology in hypertrophic cardiomyopathy. *J. Mol. Cell. Cardiol.* **157**, 77–89. <https://doi.org/10.1016/j.yjmcc.2021.04.009>.
48. Wu, H., Yang, H., Rhee, J.W., Zhang, J.Z., Lam, C.K., Sallam, K., Chang, A.C.Y., Ma, N., Lee, J., Zhang, H., et al. (2019). Modelling diastolic dysfunction in induced pluripotent stem cell-derived cardiomyocytes from hypertrophic cardiomyopathy patients. *Eur. Heart J.* **40**, 3685–3695. <https://doi.org/10.1093/eurheartj/ehz326>.
49. Fischer, T.H., Herting, J., Mason, F.E., Hartmann, N., Watanabe, S., Nikolaev, V.O., Sprenger, J.U., Fan, P., Yao, L., Popov, A.F., et al. (2015). Late INa increases diastolic SR-Ca2+-leak in atrial myocardium by activating PKA and CaMKII. *Cardiovasc. Res.* **107**, 184–196. <https://doi.org/10.1093/cvr/cvv153>.
50. Morotti, S., Edwards, A.G., McCulloch, A.D., Bers, D.M., and Grandi, E. (2014). A novel computational model of mouse myocyte electrophysiology to assess the synergy between Na<sup>+</sup> loading and CaMKII. *J. Physiol.* **592**, 1181–1197. <https://doi.org/10.1113/jphysiol.2013.266676>.
51. Hegyi, B., Borst, J.M., Bailey, L.R.J., Shen, E.Y., Lucena, A.J., Navedo, M.F., Bossuyt, J., and Bers, D.M. (2020). Hyperglycemia regulates cardiac K<sup>+</sup> channels via O-GlcNAc-CaMKII and NOX2-ROS-PKC pathways. *Basic Res. Cardiol.* **115**, 71. <https://doi.org/10.1007/s00395-020-00834-8>.
52. Ma, J., Guo, L., Fiene, S.J., Anson, B.D., Thomson, J.A., Kamp, T.J., Kolaja, K.L., Swanson, B.J., and January, C.T. (2011). High purity human-induced pluripotent stem cell-derived cardiomyocytes: electrophysiological properties of action potentials and ionic currents. *Am. J. Physiol. Heart Circ. Physiol.* **301**, H2006–H2017. <https://doi.org/10.1152/ajpheart.00694.2011>.
53. Horváth, A., Lemoine, M.D., Löser, A., Mannhardt, I., Flenner, F., Uzun, A.U., Neuber, C., Breckwoldt, K., Hansen, A., Girdauskas, E., et al. (2018). Low resting membrane potential and low inward rectifier potassium currents are not inherent features of hiPSC-derived cardiomyocytes. *Stem Cell Reports* **10**, 822–833. <https://doi.org/10.1016/j.stemcr.2018.01.012>.
54. McKeithan, W.L., Feyen, D.A.M., Bruyneel, A.A.N., Okolotowicz, K.J., Ryan, D.A., Sampson, K.J., Potet, F., Savchenko, A., Gómez-Galeno, J., Vu, M., et al. (2020). Reengineering an antiarrhythmic drug using patient hiPSC cardiomyocytes to improve therapeutic potential and reduce toxicity. *Cell Stem Cell* **27**, 813.e6–821.e6. <https://doi.org/10.1016/j.stem.2020.08.003>.
55. Itzhaki, I., Maizels, L., Huber, I., Zwi-Dantsis, L., Caspi, O., Winterstern, A., Feldman, O., Gepstein, A., Arbel, G., Hammerman, H., et al. (2011). Modelling the long QT syndrome with induced pluripotent stem cells. *Nature* **471**, 225–229. <https://doi.org/10.1038/nature09747>.
56. Prajapati, C., Ojala, M., and Aalto-Setälä, K. (2018). Divergent effects of adrenaline in human induced pluripotent stem cell-derived cardiomyocytes obtained from hypertrophic cardiomyopathy. *Dis. Model. Mech.* **11**, dmm032896. <https://doi.org/10.1242/dmm.032896>.
57. Lian, X., Hsiao, C., Wilson, G., Zhu, K., Hazeltine, L.B., Azarin, S.M., Raval, K.K., Zhang, J., Kamp, T.J., and Palecek, S.P. (2012). Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *Proc. Natl. Acad. Sci. USA.* **109**, E1848–E1857. <https://doi.org/10.1073/pnas.1200250109>.
58. Feyen, D.A.M., McKeithan, W.L., Bruyneel, A.A.N., Spiering, S., Hörmann, L., Ulmer, B., Zhang, H., Briganti, F., Schweizer, M., Hegyi, B., et al. (2020). Metabolic maturation media improve physiological function of human iPSC-derived cardiomyocytes. *Cell Rep.* **32**, 107925. <https://doi.org/10.1016/j.celrep.2020.107925>.
59. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308. <https://doi.org/10.1038/nprot.2013.143>.
60. Allaire, J.J., and Chollet, F. (2019). keras: R interface to Keras. <https://github.com/rstudio/keras>.



Science  
that inspires



## With cutting-edge research across the life sciences

Science is evolving and Cell Press is too. Our story began over 45 years ago with the journal *Cell* and a commitment to publishing exciting biology. Our journals across the spectrum of life science build on this tradition of excellence with important scientific advances.

To learn more about these titles and the other journals in our portfolio visit [cell.com](http://cell.com)

Life Science journals

**Cell** · **Med** · **Cancer Cell** · **Cell Chemical Biology** · **Cell Genomics**  
**Cell Host & Microbe** · **Cell Metabolism** · **Cell Reports** · **Cell Reports Medicine**  
**Cell Reports Methods** · **Cell Stem Cell** · **Cell Systems** · **Current Biology**  
**Developmental Cell** · **Heliyon** · **Immunity** · **iScience** · **Molecular Cell** · **Neuron**  
**Patterns** · **STAR Protocols** · **Structure**

## Report

**RECOVER identifies synergistic drug combinations *in vitro* through sequential model optimization**

Paul Bertin,<sup>1</sup> Jarrid Rector-Brooks,<sup>1</sup> Deepak Sharma,<sup>1</sup> Thomas Gaudet,<sup>2</sup> Andrew Anighoro,<sup>2</sup> Torsten Gross,<sup>2</sup> Francisco Martínez-Peña,<sup>3</sup> Eileen L. Tang,<sup>3</sup> M.S. Suraj,<sup>2</sup> Cristian Regep,<sup>2</sup> Jeremy B.R. Hayter,<sup>2</sup> Maksym Korablyov,<sup>1</sup> Nicholas Valiante,<sup>4</sup> Almer van der Sloot,<sup>5</sup> Mike Tyers,<sup>7</sup> Charles E.S. Roberts,<sup>2</sup> Michael M. Bronstein,<sup>2,6</sup> Luke L. Lairson,<sup>3</sup> Jake P. Taylor-King,<sup>2,8,9,\*</sup> and Yoshua Bengio<sup>1,8</sup>

<sup>1</sup>Mila, the Quebec AI Institute, Montreal, QC, Canada

<sup>2</sup>Relation Therapeutics, London, UK

<sup>3</sup>Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA

<sup>4</sup>Innovac Therapeutics, Inc., Cambridge, MA, USA

<sup>5</sup>IRIC, Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, QC, Canada

<sup>6</sup>Department of Computer Science, University of Oxford, Oxford, UK

<sup>7</sup>Program in Molecular Medicine, Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, 686 Bay Street, Toronto, ON M5G 0A4, Canada

<sup>8</sup>Senior author

<sup>9</sup>Lead contact

\*Correspondence: [jake@relationrx.com](mailto:jake@relationrx.com)

<https://doi.org/10.1016/j.crmeth.2023.100599>

**MOTIVATION** Galvanized by the COVID-19 pandemic, we wanted to systematically identify efficacious drug combinations from the plethora of safe drugs that could hypothetically exhibit antiviral activity. The infeasibility of extensive combinatorial screens triggered the need for new methods that would require substantially less screening than an exhaustive evaluation. Outside of biology, there has been much interest in how areas of machine learning, including active learning and sequential model optimization, can be utilized to efficiently explore large spaces of possibilities through the intelligent acquisition and interpretation of data. Sequential model optimization has received much interest within biomedicine, with a focus on systems with well-described individual components, e.g., biomolecular design, chemical assays, etc. We wanted to apply a similar philosophy to quickly identify synergistic drug combinations to alter the phenotype of a cellular model system (cell viability as proof of concept), where the relationship between the chemical inputs and resulting phenotypic output is not well understood and is subject to experimental biases.

**SUMMARY**

For large libraries of small molecules, exhaustive combinatorial chemical screens become infeasible to perform when considering a range of disease models, assay conditions, and dose ranges. Deep learning models have achieved state-of-the-art results *in silico* for the prediction of synergy scores. However, databases of drug combinations are biased toward synergistic agents and results do not generalize out of distribution. During 5 rounds of experimentation, we employ sequential model optimization with a deep learning model to select drug combinations increasingly enriched for synergism and active against a cancer cell line—evaluating only ~5% of the total search space. Moreover, we find that learned drug embeddings (using structural information) begin to reflect biological mechanisms. *In silico* benchmarking suggests search queries are ~5–10× enriched for highly synergistic drug combinations by using sequential rounds of evaluation when compared with random selection or ~3× when using a pretrained model.



## INTRODUCTION

Drug combinations are an important therapeutic strategy for treating diseases that are subject to evolutionary dynamics, in particular cancers and infectious diseases.<sup>1,2</sup> Conceptually, as tumors or pathogens are subject to change over time, they may develop resistance to a single agent<sup>3</sup>—motivating one to target multiple biological mechanisms simultaneously.<sup>4</sup> Discovering synergistic drug combinations is a key step toward developing robust therapies, as they hold the potential for greater efficacy while reducing dose and hopefully limiting the likelihood of adverse effects. For example, in a drug repurposing scenario (i.e., uncovering new indications for known drugs), the ReFRAME library of ~12,000 clinical-stage compounds<sup>5</sup> leads to ~ 72 million pairwise combinations; this does not appear tractable with standard high-throughput screening (HTS) technology—even at a single dose.<sup>6</sup> Moreover, with patient-derived organoids (PDOs) being examined as a biomarker within personalized medicine clinical studies,<sup>7,8</sup> the search space expands further to identify efficacious drug combinations specific to the mutation profile in question.

With the recent COVID-19 global health crisis, there has been the need for rapid drug repurposing that would allow for expedited and derisked clinical trials. Due to the complexity of selecting drug combinations and the minimal training data publicly available, studies have typically been limited toward monotherapy repurposing from a variety of angles—often involving artificial intelligence (AI) techniques to provide recommendations.<sup>9</sup> The dearth of drug combination datasets is due to the large combinatorial space of possible experiments available—ultimately limiting the quality of drug synergy predictions. Moreover, databases of drug combinations are biased toward suspected synergistic agents, and thus making predictions outside the scope of the training dataset can be challenging.

The goal of this work is to discover synergistic drug combinations while only requiring minimal wet-lab experimentation. One cost-efficient tool at our disposal is sequential model optimization (SMO), whereby a machine learning (ML) model selects experiments (i.e., pairs of drugs) that it would like to be evaluated (in this case, for drug synergism). Both highly informative experiments (“exploration”) and experiments that double down on promising data-driven hypotheses (“exploitation”) can be selected.<sup>10</sup> Between rounds of experimental evaluation, the model is iteratively adapted to new observations (via model training), which allows performance to gradually improve. This SMO process allows for queries that are more and more enriched with highly synergistic combinations, ultimately leading to reduced experimentation when compared to an exhaustive search.

There have now been a number of approaches for predicting the effects of and subsequently prioritizing drug combinations.<sup>11</sup> Classic bioinformatics approaches have focused on using ML and network statistics over specified features of drugs (e.g., molecular fingerprints<sup>12</sup>), cell lines (e.g., transcriptomics, copy-number variations<sup>13</sup>), and interactome topology between biomolecules (e.g., protein-protein interactions, chemical-genetic interactions,<sup>14</sup> or gene regulatory networks<sup>15</sup>). Initiatives such as the Dialogue on Reverse Engineering Assessment and

Methods (DREAM) have led to a plethora of methods being benchmarked against one another in prospective challenges through the generation of novel datasets.<sup>16</sup> Complex deep learning architectures, which have set state-of-the-art performance across a number of domains,<sup>17</sup> have been used to predict both adverse drug-drug interactions<sup>18,19</sup> and synergistic drug combinations.<sup>20–22</sup> Sequential approaches, wherein several rounds of selection are performed, have also been explored in the context of drug combinations; for example, Kashif et al.<sup>23</sup> have proposed a heuristic-based (as opposed to a model-based) exploration strategy.

We present a SMO platform that can guide wet-lab experiments: RECOVER, a deep learning regression model that predicts synergy using molecular fingerprints as inputs. To motivate the use of RECOVER, we demonstrate a real-world use case whereby one observes both: a ~5–10 $\times$  estimate for the enrichment of synergistic drugs identified using SMO when compared with selecting drug combinations at random and a ~3 $\times$  improvement when compared with selecting drugs in a single batch using a pretrained model. We then perform a retrospective validation to benchmark the performance of our model and understand its generalization abilities using the DrugComb database—largely pertaining to cancer cell line data.<sup>24</sup> Thereafter, we evaluate our SMO pipeline *in silico*, which allows the model to select the most relevant data points to be labeled in order to discover the most promising combinations while reducing model uncertainty. Finally, we test RECOVER prospectively in an *in vitro* experimental setting, whereby we discover novel synergistic combinations active against a breast cancer model cell line, MCF7, which is also represented within our training dataset.

With an SMO platform available in conjunction with an appropriate *in vitro* assay, one has a powerful tool to rapidly respond to a future public health crisis. To encourage use by the scientific community, we detail a configuration that can be trained on a personal computer or laptop without requiring dedicated computational infrastructure. Remarkably, high predictive power is not a prerequisite for such an SMO system to be utilized effectively. In fact, as we are trying to identify pairs of drugs in prospective experiments that have more extreme synergy scores than those drug combinations evaluated within previous experiments (i.e., our training dataset), we cannot necessarily expect to have high predictive power. However, we achieve our ultimate goal: the identification of highly synergistic drugs—not building highly accurate ML models. This work forms a proof-of-concept demonstration of RECOVER—which should then motivate greater community adoption of the method and extensions thereof.

## RESULTS

### RECOVER: SMO platform for rapid drug repurposing

RECOVER is an open-source SMO platform for the optimal suggestion of drug combinations (see Figure 1). Pairs of drug feature vectors are fed into a deep neural network, which is used for the prediction of synergy scores. These feature vectors include molecular fingerprints as well as a one-hot encoding identifying a drug. For a full description of the model, see method details and Figure S4A.

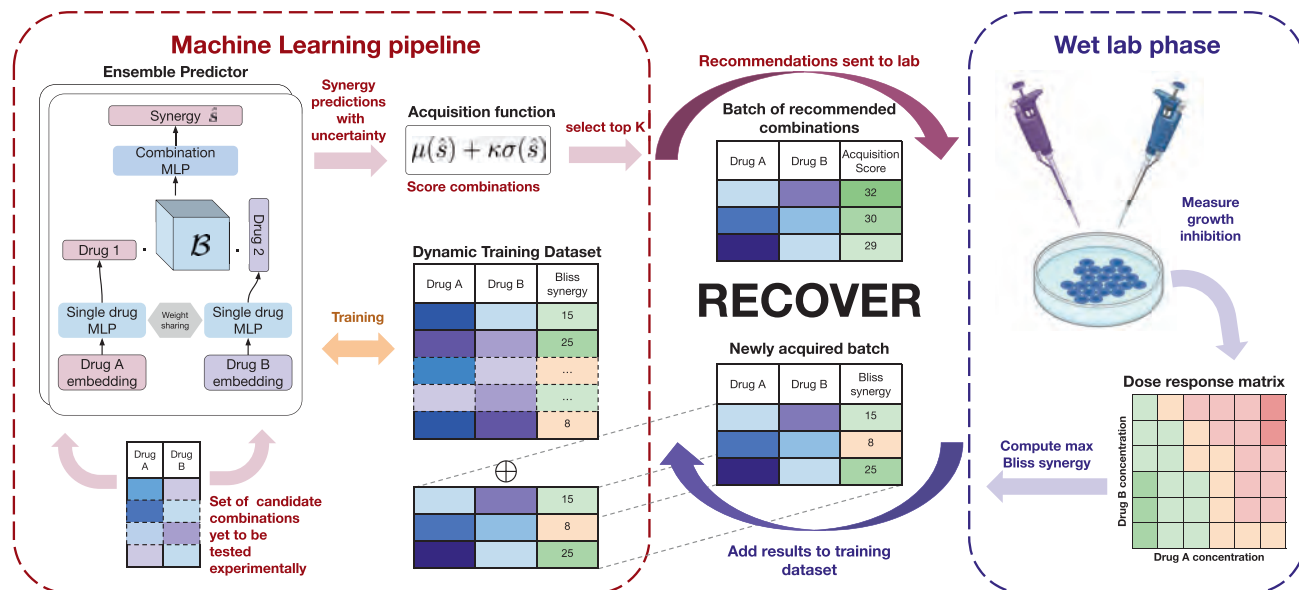


Figure 1. Overview of the RECOVER workflow integrating both a novel machine-learning pipeline and iterated wet-lab evaluation

Our core focus is the prediction of pairwise drug combination synergy scores. While many mathematical descriptions of synergy have been proposed,<sup>1</sup> in the following work, we utilize the Bliss synergy score due to its simplicity and numerical stability. In the context of cell viability, the Bliss independence model assumes that in the absence of synergistic effects, the expected fraction of viable cells after treatment with drugs  $d_1$  and  $d_2$  at doses  $c_1$  and  $c_2$ , written  $V(c_1, c_2)$ , is identical to the product of the fractions of viable cells when utilizing each drug independently, i.e.,  $V(c_1)V(c_2)$ . We then define the Bliss synergy score as the difference between these quantities such that a fraction of surviving cells  $V(c_1, c_2)$  smaller than the expected proportion  $V(c_1)V(c_2)$  leads to a large Bliss synergy score,

$$\begin{aligned} S_{Bliss}(c_1, c_2) &= V(c_1)V(c_2) - V(c_1, c_2) \\ &= I(c_1, c_2) - I(c_1) - I(c_2) + I(c_1)I(c_2), \end{aligned} \quad (\text{Equation 1})$$

where  $I(\cdot) = 1 - V(\cdot)$  is the experimentally measured growth inhibition induced by drug  $d_1$ ,  $d_2$ , or both together at the associated doses. Given a dose-response matrix for the two drugs, a global synergy score can be obtained through a pooling strategy. In our case, we take the maximum value, i.e.,

$$\hat{S}_{Bliss} = \max_{c_1, c_2} S_{Bliss}(c_1, c_2). \quad (\text{Equation 2})$$

In many studies, the arithmetic mean is taken to calculate a global synergy score. Unfortunately, different laboratories use different dose intervals for each drug, and typically, each drug combination shows a synergistic effect at a specific dose-pair interval. Therefore, the arithmetic mean is highly sensitive to the chosen dose interval and is thus why we choose to prioritize a max-pooling strategy as in Equation 2. Unless explicitly stated otherwise, a synergy score refers to a global max-pooled Bliss score.

In addition to the prediction of synergy, RECOVER estimates the uncertainty associated with the underlying prediction. More precisely, for a given combination of drugs, RECOVER not only provides a point estimate of the synergy but estimates the distribution of possible synergy scores for each combination, which we refer to as the predictive distribution. We define the model uncertainty as the standard deviation of the predictive distribution.

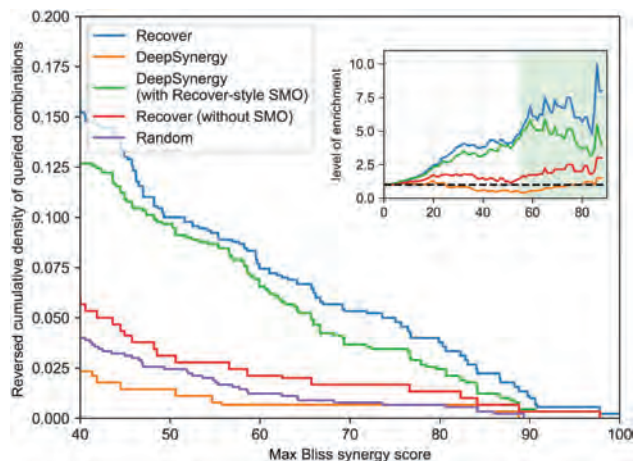
An acquisition function is used to select the combinations that should be tested in subsequent experiments.<sup>25</sup> This acquisition function is designed to balance between *exploration*, prioritizing combinations with high model uncertainty, whereby labeling said points should increase predictive accuracy in future experimental rounds; and *exploitation*, the selection of combinations believed to be synergistic with high confidence.

In summary, this SMO setting consists of generating recommendations of drug combinations that will be tested *in vitro* at regular intervals. At each step, RECOVER is trained on all the data acquired up to that point, and predictions are made for all combinations that could be hypothetically tested experimentally. The acquisition function is then used to provide recommendations for *in vitro* testing. The results of the experiments are then added to the training data for the next round of experiments, and the whole process repeats itself.

#### Task variations

We note that there are two separate but related frameworks in which RECOVER can be utilized.

In the preclinical framework, RECOVER can be used to recommend drug combinations expected to be effective within a single specified cell model system: the model is asked to provide synergy predictions from inputs  $(d_1, d_2)$  for drugs  $d_1$  and  $d_2$  and to subsequently provide recommendations in the same format. The preclinical framework is most relevant to early drug discovery; for example, one may wish to prioritize assets within a portfolio that synergize with an already approved drug. Naturally, we



**Figure 2. Simulations suggest that RECOVER can enrich for highly synergistic combinations given a limited budget**

Reversed cumulative density of queried combinations following different querying strategies. (Inset) Level of enrichment. Shaded area corresponds to synergies > 54.9. Results are averaged over 3 seeds.

can apply RECOVER to any disease areas where *in vitro* cell models are routinely used in early drug discovery, e.g., collagen deposition (fibrosis), T cell activation (immunology), etc.

In an alternative setup, the personalized framework requires RECOVER to recommend drug combinations expected to be effective in one or more available model systems: the model is asked to provide predictions and subsequent recommendations of the form  $(d_1, d_2, m)$  for drugs  $d_1, d_2$ , and model system  $m$ . The personalized framework is most relevant to novel personalized cancer treatment scenarios, wherein multiple patient-derived primary models are available and recommendations are sought to optimize the use of approved drugs in a highly translatable but low-throughput system.<sup>26,27</sup>

### Illustration of SMO approach

To illustrate the benefits of the SMO approach, we perform a preliminary simulation to mimic a scientist with a limited experimental budget of 300 drug combinations to be tested—with the aim to find synergistic drug combinations. We assume that the experimentalist has access to a trained ML model, and we show the benefit of RECOVER within both frameworks. At a high level, we specify that there are two options: either to perform all 300 experiments in one go, or to perform experiments in 10 batches of 30.

We note that many ML papers focus on the personalized framework,<sup>20,28–30</sup> i.e., recommendations are of the form  $(d_1, d_2, m)$ , so we demonstrate the benefit of SMO in this scenario first. All models are pretrained on the O’Neil drug combination study,<sup>31</sup> and validation by the experimentalist is simulated through uncovering specific examples from the NCI-ALMANAC drug combination study<sup>32</sup> restricted to all cell lines that are covered in both studies. In more detail, we test the following options: random, all 300 combinations are queried at random; DeepSynergy, the synergies of all combinations in ALMANAC are predicted using the DeepSynergy model with the top 300 predictions queried; RECOVER without SMO, the synergies of

all combinations in ALMANAC are predicted using the RECOVER model with the top 300 predictions queried; RECOVER, 30 combinations are queried at random followed by an SMO using batches of 30; and DeepSynergy with SMO, which is the same SMO as before but using the DeepSynergy model.

In Figure 2, we report the reversed cumulative density of the synergies of all 300 queried combinations (higher is better). We also report the level of enrichment defined as the ratio between the reversed cumulative density of a given strategy’s queries and the reversed cumulative density of random queries. We first observe that DeepSynergy<sup>20</sup> performs worse than random, while RECOVER (without SMO) performs slightly above the level of randomness. Most importantly, the bulk of the performance gain comes from utilizing our SMO procedure. Finally, when RECOVER and DeepSynergy are compared head to head in the SMO setting, the RECOVER model outperforms the DeepSynergy model.

The threshold for “highly synergistic” is challenging to specify, but we note that a drug combination in clinical trials has a max Bliss synergy score of 54.9 (see discovery and rediscovery of novel synergistic drug combinations). On this basis, these experiments suggest that our approach can reduce by a factor of ~5–10× the number of experiments needed to discover and validate highly synergistic drug combinations when compared with random selection or by a factor of >3× when using a pretrained model selecting all drug combinations at a single time point.

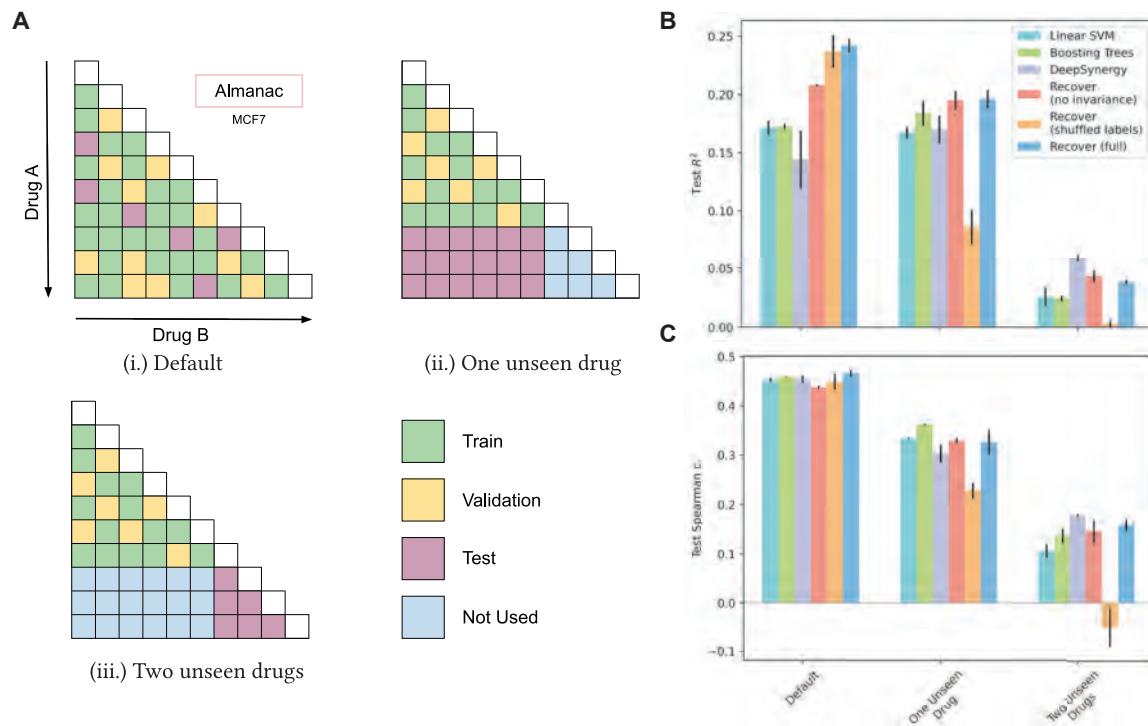
For completeness, we show in Figure S1A that we achieve a broadly similar level of enrichment when evaluating a preclinical framework task for three different cell lines. The experimental setup is exactly the same except that the search space is now restricted to a specific cell line within the NCI-ALMANAC study and recommendations are of the form  $(d_1, d_2)$ . We note that tasks drawn from the preclinical framework are slightly more challenging than the tasks drawn from the personalized framework, as the model cannot evaluate the same drug pairs in new cell lines (which would likely lead to drug synergy), and so the performance is marginally lower.

### Scope of RECOVER capabilities and experimental validation

Due to the operational complexities in prospectively evaluating performance in the personalized framework, we focus on the preclinical framework for experimental proof of concept and demonstration of the RECOVER system. In Figure S1H, we report key aspects of our prospective validation and how it compares with the ones performed in other published works. We note that other works focused on generalizing to a new cell line and/or combinations of drugs both seen during training. Our prospective validation focuses on testing the ability of RECOVER to generalize to combinations involving one drug seen during training and one unseen drug, which is a harder task. In addition, validation involves, for the first time, repeated experimentation via an integrated wet-lab/dry-lab system.

### Retrospective testing of RECOVER informs the design of future experiments

In preparation for prospective validation within the preclinical framework, we evaluate the performance of RECOVER *in silico*



**Figure 3. Retrospective testing demonstrates the ability of RECOVER to generalize when at least one of the drugs has been seen during training but not beyond that**

(A) Overview of the different tasks on which RECOVER has been evaluated in preparation for the prospective evaluation within the preclinical framework. Each task corresponds to a different way to split the training, validation, and test sets and aims at evaluating a specific generalization ability of the model. (i.) Default. Combinations are split randomly into training/validation/test (70%/20%/10%). Only the MCF7 cell line is used. (ii.) One unseen drug. 30% of available drugs are excluded from the training and validation sets. The test set consists of combinations between a drug seen during training and an unseen drug. Combinations among seen drugs are split into training and validation (80%/20%). Only the MCF7 cell line is used. (iii.) Two unseen drugs. Similar to task (ii.), but the test set consists of combinations of two unseen drugs.

(B and C) Performance of RECOVER and other models for the three different tasks. Standard deviation computed over 3 seeds.

using previously published data. In order to understand the scope of scenarios to which RECOVER can be applied to, we benchmark RECOVER against baseline models and test our ability to generalize in several out-of-distribution tasks without incorporating SMO. Thereafter, we perform backtesting through simulating mock SMO experiments (see SMO development and evaluation in the method details, as well as in Figures S4D–S4F).

Due to the limited size of most individual drug combination studies reported in the literature, we focus on the NCI-ALMANAC viability screen<sup>32</sup> summarized in Figure S1B. We refrain from combining multiple datasets because of the severe batch effects between studies; in Figure S1F, we show a scatterplot that demonstrates inconsistency between the O’Neil et al.<sup>31</sup> series of drug combination experiments against their NCI-ALMANAC counterpart. We note that this may result from variation in the readouts of these experiments, mutations in cell lines, or differences in harvest times.

We investigate whether RECOVER can generalize beyond the training (and validation) set in various ways: (Figure 3Ai.) what is the performance on test cases drawn from the same distribution as the training set? Can RECOVER generalize when (Figure 3Aii.) one of the drugs is unseen (during training) or (Figure 3Aiii.) when

both of the drugs are unseen? These tasks are illustrated graphically in Figure 3A. For each task, we benchmark against several alternative models along with RECOVER, including a linear support vector machine (SVM), Boosting Trees, and DeepSynergy.<sup>20</sup> In addition, we evaluate a version of RECOVER without the invariance module and another version for which the identities of the drugs (as well as cell lines) have been shuffled (see model development and evaluation in the method details for further information on models and hyperparameter optimization procedures). Through understanding the capability of RECOVER to generalize, we can design prospective experiments with a greater confidence of success.

In Figures 3B and 3C, we report the test performance metrics of RECOVER across each of the first three tasks. Examining performance within task (i.) in Figure 3A, test statistics appear modest; however, we demonstrate limits on achievable performance—resulting from experimental noise and non-uniformity of synergy scores (see Figure S2F). From task (i.) to task (iii.) in Figure 3A, we note a drastic drop in performance for all models, but this effect is alleviated if only one of the drugs has not been seen before (see task ii. in Figure 3A). We also investigate additional scenarios from the personalized framework, presented in Figure S2A, wherein we consider multiple cell lines, as well as

training and test sets coming from different studies, and report performance in Figure S2B.

We note that our benchmarking justifies various aspects of our deep learning architecture: the RECOVER permutation invariance module can provide improvement in performance across some scenarios; moreover, RECOVER (shuffled labels) fails compared with other methods on task (ii.) in Figure 3A with one unseen drug and is at the level of randomness on task (iii.) in Figure 3A with two unseen drugs. In these cases, we demonstrate that drug structure is actually leveraged by the model in order to generalize (to some extent) to unseen drugs. However, RECOVER (shuffled labels) performs well compared with other models on the default task; thus, merely knowing the identity of the drugs is sufficient when both drugs have been seen in other combinations.

From the above results, we can recommend that any prospective experiments should require that one of the two drugs in the combination have been seen in some context before (see task iii. in Figure 3A). Due to the severe batch experiments between studies in the public domain, as shown in Figure S1F, models fail to generalize to data coming from a different study, as shown in Figure S2B (study transfer task). As such, should we want to utilize publicly available resources, we will have to incorporate such data intelligently. To this end, we investigated using transfer learning, wherein one trains a model on a large dataset (known as pretraining) and thereafter refines the model on a smaller dataset (known as fine-tuning)—typically with some aspect of the task or the data changed between the two instances. We show that this is possible and beneficial (compared to not leveraging existing data) in an SMO setting between the O’Neil et al.<sup>31</sup> and NCI-ALMANAC studies (see Figure S4E). Remarkably, even with minimal correlation between studies, we are able to observe the benefits of transfer learning in this scenario. These findings suggest that we use transfer learning within prospective experiments.

### Prospective use of RECOVER enriches for selection of synergistic drug combinations

From the *in silico* results, we now test RECOVER prospectively using a cancer cell model, leveraging publicly available data for pretraining. Using the insights from retrospective testing of RECOVER informs the design of future experiments, the queryable space of drug combinations was designed to include drug pairs where only one compound was already seen by the model during pretraining—with a second compound not seen before. For details about the model used to generate recommendations, see recommendation generation in the method details. The MCF7 cell line was used to generate 6×6 dose-response matrices (see experimental protocol for details).

We perform multiple rounds of RECOVER-informed wet-lab experiments and observe sequential improvements in performance. The rounds of experiments are described as follows.

- (1) Calibration. The initial round of experiments was performed to supplement publicly available data with 20 randomly selected unseen drug combinations. Furthermore, we confirmed the previous *in silico* result that we could not predict synergy scores (prior to transfer learning adaptation) through selecting 5 highly synergistic combi-

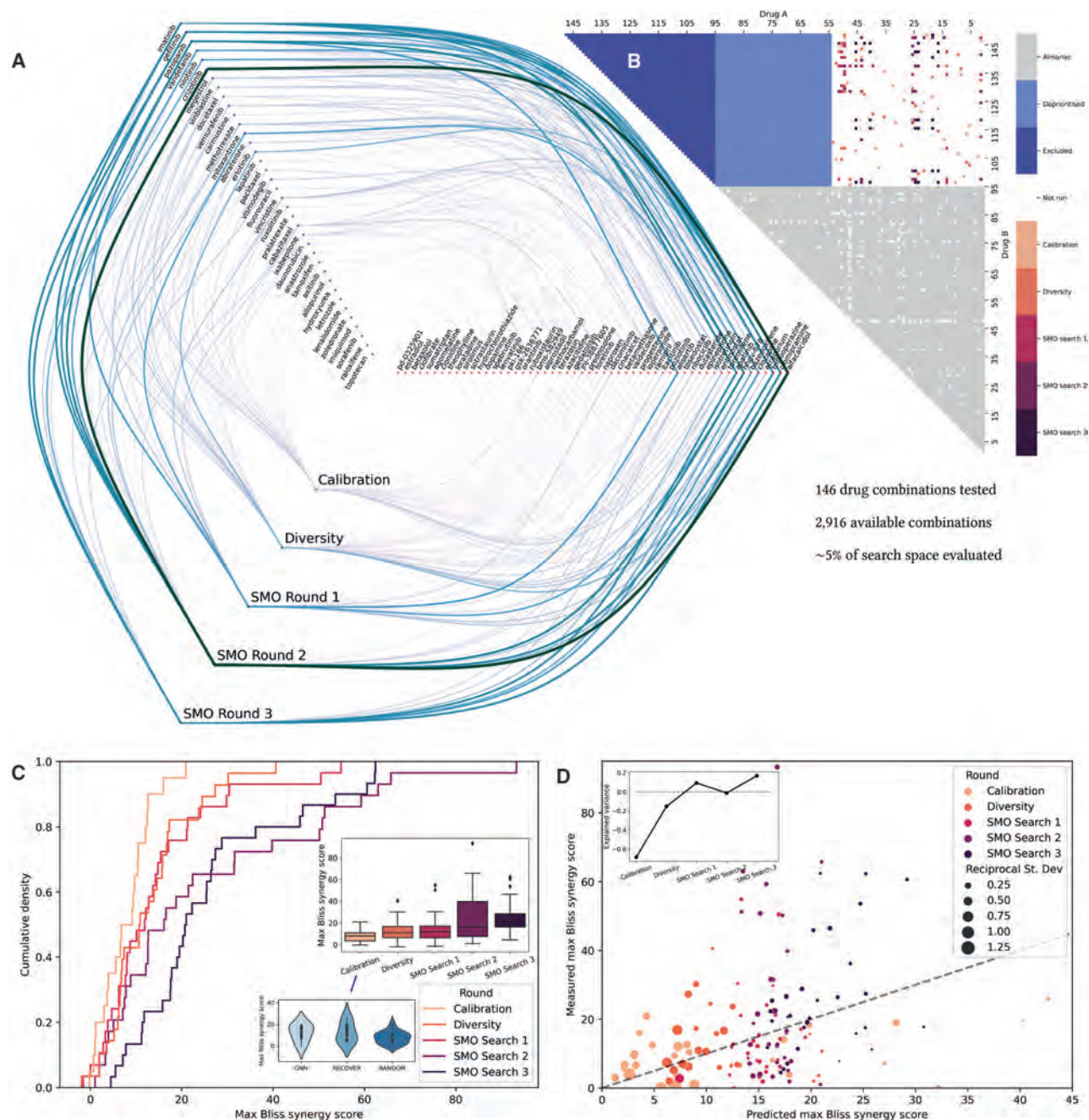
nations selected by RECOVER. In addition, 5 more drug combinations were selected by a graph neural network (GNN) model in the style of Zitnik et al.<sup>18</sup> that we did not develop further due to the computational overhead. It was also specified that each drug should appear in, at most, a single drug combination queried.

- (2) Diversity. Thereafter, drug combinations are selected using model predictions in conjunction with the upper confidence bound (UCB) acquisition function. To ensure that we quickly observe all single drugs at least once (as we showed that the model cannot generalize well to combinations involving unseen drugs), we select our batch of experiments as follows. First, we rank combinations according to their acquisition function score. We then find the first combination that involves a drug that has not yet been used (or that is involved in one of the combinations from the current batch) and add it to the batch. We repeat this until we have 30 combinations in the batch.
- (3) SMO search. RECOVER is now free to select any drug pairs of interest for testing, with the requirement that any single drug may be selected no more than 5 times (to avoid oversampling and depletion of chemical stock). Three such rounds have been performed in this manner.

The search space was constructed as follows. The NCI-ALMANAC includes 95 unique drugs that were employed in combinations tested on the MCF7 cell line (see gray area in Figure 4B). We chose to deprioritize drugs without a well-characterized mechanism of action (MoA) to facilitate biological interpretation and validation of the results (see light blue area in Figure 4B). To achieve this, drugs in NCI-ALMANAC were annotated with known targets extracted from the ChEMBL drug mechanism table: 54 drugs matched with at least one known target were thus selected. An additional 54 drugs were selected by clustering drugs with known MoAs that are included in the DrugComb<sup>24</sup> database but not in NCI-ALMANAC. Hence, a search space including a total of 2,916 drug combinations was obtained (see the white area in Figure 4B). In Figure 4A, we illustrate the pairs of drugs selected in each round of experiments.

We now evaluate both the synergy scores of the drug combinations selected and the underlying accuracy of the model. In Figure 4C, we plot the cumulative density function of each experimental round. We note that the mean of the max Bliss synergy scores significantly increases between the first and the third rounds (t test,  $p < 0.05$ ); this trend further continues by the fifth round (t test,  $p < 10^{-5}$ ). Moreover, the distribution starts developing a heavier tail toward high max Bliss synergy scores. This emergent heavy tail already appears significant when comparing the distribution in the first SMO search round to the background distribution of synergy scores in NCI-ALMANAC (Kolmogorov-Smirnov test,  $p < 0.025$ ). Finally, the highest max Bliss synergy score observed increases between rounds until the second SMO search round, whereby the behavior appears to have saturated. These results are focused on the max Bliss score, which RECOVER was specifically designed to optimize for; for completeness, we also report similar evaluations based on different aggregation strategies of the Bliss scores (see Figure S3A).





**Figure 4. *In vitro* evaluation demonstrates the significant enrichment for highly synergistic combinations through prospective use of RECOVER**

(A) Network plot indicating which pairs of drugs were identified at each round; line color and width represent synergy.

(B) Heatmap representing drug combinations used during pretraining (NCI-ALMANAC), in the five subsequent rounds of experiments, and combinations excluded from the analysis. Drug combinations that were not available for pretraining or were not selected for experiments are represented in white.

(C) Cumulative density plot of max Bliss synergy score for each experimental round; (inset) boxplot representation and calibration round details.

(D) Predicted versus actual plot for max Bliss synergy score. The dotted line corresponds to  $y = x$ . (Inset) The explained variance is plotted for each experimental round.

See also Data S1 and Tables S1 and S2.

All combinations queried throughout the five rounds, and their corresponding synergy scores, are provided in Table S1. We notice that specific drugs tend to appear in several of the combinations recommended by RECOVER. Consistent with the literature, we observe that some compounds appear more often than others within synergistic combinations,<sup>33</sup> a pattern that can also be observed within the NCI-ALMANAC study (see Figure S1C). However, this does not make the identification of synergistic combinations a trivial problem: even drugs that lead to the highest number of synergistic combinations are non-synergistic most of the time. No single drug within the NCI-ALMANAC study has a synergy score >40 more than 10% of the time (or 12% when considering only the MCF7 cell line data within the NCI-ALMANAC study; see Figure S1G). In comparison, our last two rounds of *in vitro* experiments yielded 20%–30% of combinations with a synergy >40 (see Figure 4C), while the model had only observed less than 5% of the search space.

In Figure 4D, we plot the predicted versus actual plot of the max Bliss synergy score. Here, the point size in the scatterplot is inversely proportional to the model uncertainty; therefore, we display confident predictions as large points, and vice versa. As expected, more confident predictions are closer to the  $y = x$  line. Less-confident predictions are associated with larger max Bliss synergy scores. Moreover, we systematically underestimate the measured max Bliss synergy score (more points far above  $y = x$  line); this intuitively makes sense, as we are trying to identify highly synergistic drug combinations that are not within our training dataset. Figure 4D (inset) displays the increase in (weighted) explained variance from one round to the next; weights are chosen to be the reciprocal of the model uncertainty. We find that, initially, the explained variance is negative, i.e., our model has no predictive power. However, as the experiments continue, a positive trend emerges such that we have a small amount of predictive power by the end of the experiments.

This increase in performance and in the synergy of queried combinations from one round to the next demonstrated in Figure 4C is expected and can be attributed to two factors. First, we needed to adapt the model to predict in a new experimental setting. From the study transfer task in Figure S2A, we know that this would otherwise be an impossible task and thus motivates the calibration round. After the calibration round, one expects that the systematic biases learned by the model during pretraining are minimized. At this point, the model is in a scenario akin to task (ii.) in Figure 3A. Second, we can improve performance further by enforcing that (almost) all drugs have been evaluated at some point, which subsequently motivated the diversity round. Thereafter, the model is free to optimize during the SMO rounds to the extent that it is able to, leveraging model predictions and model uncertainties. In fact, due to activity cliff effects,<sup>34</sup> there are likely fundamental limits on quantifying the relationship between model uncertainty and model error; in Figures S4B and S4C, we perform a preliminary investigation of these relationships. From our prospective use of RECOVER, we not only discover highly synergistic drug combinations but also demonstrate that high predictive power is not strictly necessary to identify synergistic drug combinations.

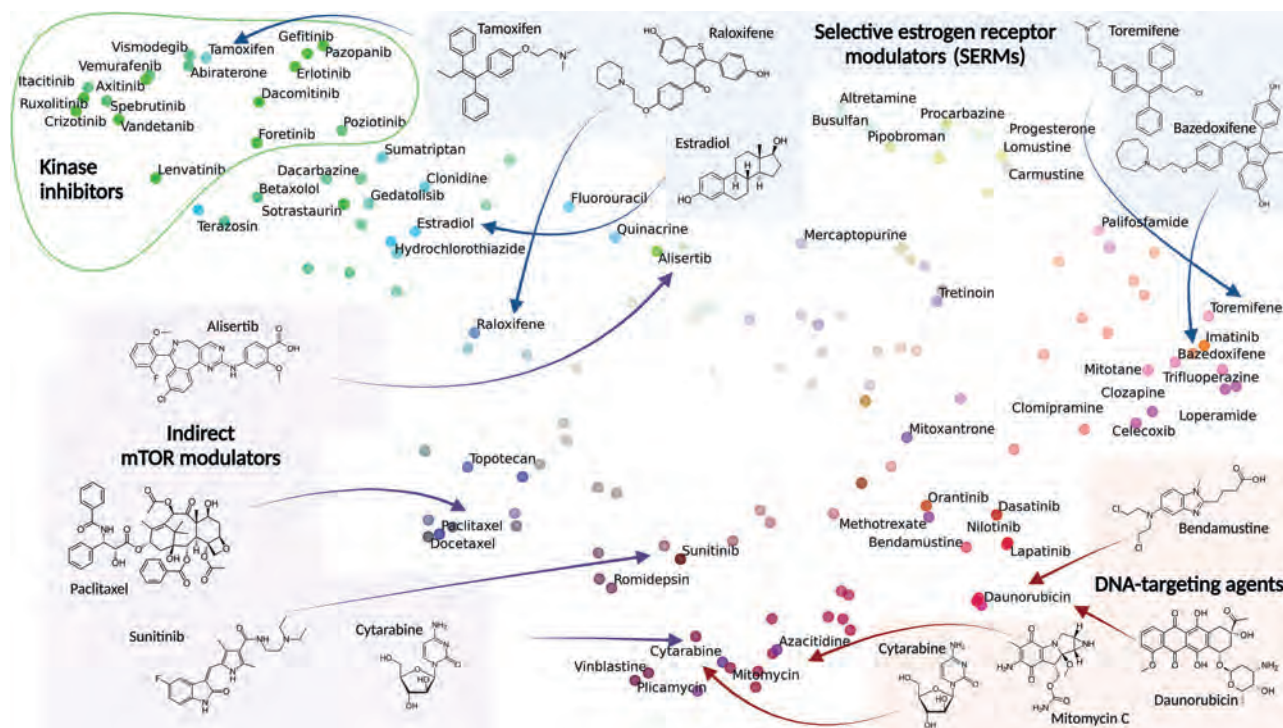
### Discovery and rediscovery of novel synergistic drug combinations

In Data S1, we provide detailed information on our experimental results using the Combenefit package<sup>35</sup> (including single-agent dose-response curves, combination dose-response surfaces, and synergy levels) for the 14 most synergistic drug combinations (from the ~150 tested), with alfacalcidol and crizotinib achieving a max Bliss score above 90. Of note, we rapidly discover drug combinations with similar mechanisms and efficacy to those already in clinical trials. Namely, within the first SMO search round we found (1) alisertib and pazopanib and (2) flumatinib and mitoxantrone. The concentration intervals for the drugs used in both drug combinations that show synergy are consistent with therapeutically relevant plasma concentrations<sup>36,37</sup> or as observed in *in vivo* animal experiments (flumatinib).<sup>38</sup>

Pazopanib inhibits angiogenesis through targeting a range of kinases including vascular endothelial growth factor receptor (VEGFR), platelet-derived growth factor receptor (PDGFR), c-KIT, and fibroblast growth factor receptors (FGFRs); in contrast, alisertib is a highly selective inhibitor of mitotic Aurora A kinase. Synergism between the two agents is hypothesized to be linked to the observation that mitosis-targeting agents also demonstrate antiangiogenic effects. In an independent study, the combination of alisertib and pazopanib has successfully completed phase 1b clinical trials for advanced solid tumors.<sup>36</sup> The combination of flumatinib and mitoxantrone appears to be linked to a similar mechanism but does not seem to have been studied in the biomedical literature. While flumatinib is a tyrosine kinase inhibitor targeting Bcr-Abl, PDGFR, and c-KIT, mitoxantrone is a type II topoisomerase inhibitor.

RECOVER drug embeddings capture both structural and biological information. To get a better insight into the drug embeddings learned by RECOVER, we report uniform manifold approximation and projection (UMAP) visualizations of the drug embeddings generated by the single-drug module in Figure 5. The color of each point is chosen by applying principal-component analysis (PCA) to the binary matrix of drug-targets and scaling the first 3 dimensions into an RGB triplet; high transparency indicates drugs with a PCA target profile close to the average PCA target profile (calculated over all drugs). In short, the position of the points indicates what RECOVER has learned about the drugs, and the color represents information known about drug mechanisms from other databases not used in the training procedure.

We note that the RECOVER model does not use information on drug targets; however, drugs with similar colors are located within similar areas of UMAP space. We also observe broad sensible patterns in UMAP space based on structure; for example, most kinase inhibitors (with the *-nib* suffix) appear in the top left hand of the UMAP. Moreover, drugs with similar mechanisms tend to be co-located; for example, see structurally diverse DNA-targeting agents in the bottom right of the UMAP. As a counterpoint, we observe that agents with either mixed agonist/antagonist profiles, including selective estrogen receptor modulators (SERMs), or targeting genes through indirect mechanisms, including mammalian target of rapamycin (mTOR), lead to less structured patterns in UMAP space. We believe that this is a



**Figure 5. RECOVER tends to map molecules with common biological mechanisms closely together (reflected by the similar colors of nearby points), even when structures are dissimilar**

UMAP of RECOVER drug embeddings with the color scheme generated to indicate the known target profile of the drugs; drugs that have molecular targets in common will have similar colors. Drug embeddings are learned using information from drug structures and viability screen data only.

highly novel observation and that it suggests that were this screen to be scaled to a larger library of small molecules, one may be able to group diverse structures into common biological mechanisms.

## DISCUSSION

Drug combinations can achieve benefits unattainable by monotherapies and are routinely investigated within clinical trials (e.g., PD-1/PD-L1 inhibitors combined with other agents<sup>39</sup>) and utilized within clinical practice (e.g., antiretroviral treatment of HIV where between 3 and 4 agents may be used<sup>40</sup>). To this end, we have presented the SMO toolbox RECOVER for drug combination identification. We have demonstrated its ability to generalize to combinations involving one unseen drug, and crucially, we have shown the benefit of repeated experimentation via an integrated wet-lab/dry-lab system. We showcase a general methodology, consisting of careful analysis of the properties of our ML pipeline—such as its out-of-distribution generalization capacities—to help us design key aspects of our prospective experiments, to eventually ensure a smooth and successful interaction between the SMO pipeline and the wet lab. Highly synergistic drug combinations have been identified, and the resulting learned drug embeddings appear to capture both structural and biological information. RECOVER can quickly (in our prospective experiments: <5% of the total search space evaluated) identify patterns in the drug-drug landscape of synergies, in or-

der to provide recommendations significantly enriched for synergism and alleviate the need for exhaustive studies. We provide commentary on key aspects on our approach covering datasets, computational methodology, wet-lab techniques, and evaluation metrics.

We note the considerable difficulties of working with publicly available datasets with discrepancies in the data generation process. Inconsistent media between multiple labs, the presence of *de novo* mutations within immortalized *in vitro* cell models, and differences in experimental protocols limit ease of data integration between laboratories.<sup>41</sup> In particular, systematic biases limit generalizability of model predictions to subsequent prospective experiments. Within oncology, protein-coding mutations may drive resistance to any one chemotherapeutic agent but also large-scale gene dosing changes from non-coding mutations,<sup>42</sup> copy-number variations,<sup>43</sup> and aneuploidy.<sup>44</sup> These issues have been somewhat alleviated through careful choice of metric to optimize (e.g., max pooled Bliss synergy scores have reduced sensitivity to selected drug concentration ranges, compared to averaged scores) and only using publicly available data for pre-training (when compared with using these data for prediction without adaptation).

From a computational perspective, we experimented with a range of more complicated models. For example, we considered using GNNs to model biomolecular interactions,<sup>45</sup> which have numerous benefits including greater biological interpretability and incorporation of prior knowledge, namely drug-target and

protein-protein interactions. However, these models only resulted in marginal increases in performance while requiring substantially more computational resources. We believe that the limited diversity of the dataset and the simplicity of the task, a one-dimensional regression, did not allow these more advanced approaches to reach their full potential. Therefore, we prioritized a strategy that could be run quickly for rapid turnaround of recommendations for experimental testing.

When considering an SMO setting, we are required to collapse highly complex information into a single number to be optimized (i.e., a synergy score). While there is an opportunity to improve choices of metric (synergy scores may not reflect absolute cell viability), assay readouts that better characterize cell state (compared with cell viability) may provide a stronger starting point. In particular, an omics readout, through transcriptomics<sup>46</sup> and/or single-cell profiling,<sup>47,48</sup> and high content imaging<sup>49</sup> provide a much higher-dimensional measurement of cell state. Furthermore, derived properties from these readouts may be more interpretable, e.g., pathway activation<sup>50</sup> or extracellular signaling.<sup>51</sup> Remarkably, even while only using cell viability as a readout, we achieved significant progress in identifying novel synergistic drug combinations.

Furthermore, the usual metrics for the evaluation and training of regression models may not reflect well the efficiency of models in iterative settings. This is due to the fact that, in our SMO setting, only the prediction of extreme values is important. This work provides an example of this effect: model performance on prospectively queried combinations was modest, but a substantial enrichment was achieved. Some metrics have been proposed to focus specifically on the prediction of extreme values.<sup>52</sup> Developing training objectives that specifically aim at maximizing SMO performance will be the object of future work.

From the systematic screen by Jaak et al.,<sup>33</sup> they conclude that synergy between drugs is rare and highly context dependent. RECOVER provides a means to identify such synergies while requiring substantially less screening than an exhaustive evaluation; thus, we expect that RECOVER and similar such systems may have a role to play when addressing diverse application areas such as personalized cancer treatment and novel emergent infectious disease such as the COVID-19 pandemic.

### Limitations of the study

In addition to the points mentioned above, a few restrictions were necessary in the name of feasibility concerning the validation experiments. In particular, only one cell model was used for validation, and the exhaustive evaluation of every possible drug combination was not performed. With regard to the downstream analysis, while we investigated the relationship between drugs and their mechanisms of action, many such mechanisms are not fully elucidated. Finally, our investigation into the relationship between the structural similarity of drug pairs, their synergy, the associated model error, and model uncertainty is preliminary in nature.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Model description
  - Searching the space of drug combinations
  - Recommendation generation
  - Dataset processing
  - Experimental protocol
  - Combenefit preprocessing
  - Model development & evaluation, excluding SMO
  - SMO development and evaluation (*in silico*)
- QUANTIFICATION AND STATISTICAL ANALYSIS

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.crmeth.2023.100599>.

### ACKNOWLEDGMENTS

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation (INV-019229). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the author-accepted manuscript version that might arise from this submission. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. The authors would also like to thank Andrew Trister, Isabelle Lacroix, David Roblin, Benjamin Swerner, Jyothish Soman, and Lindsay Edwards for useful discussion and support.

### AUTHOR CONTRIBUTIONS

Conceptualization, C.E.S.R. and J.P.T.-K.; methodology, P.B., J.R.-B., D.S., T. Gaudelet, M.S.S., M.K., M.M.B., J.P.T.-K., and Y.B.; software, P.B., J.R.-B., D.S., and T. Gaudelet; validation, F.M.-P., E.L.T., J.B.R.H., N.V., and L.L.L.; formal analysis, P.B., A.A., and T. Gross; investigation, F.M.-P. and E.L.T.; resources, J.B.R.H. and L.L.L.; data curation, A.A., T. Gross, and C.R.; writing – original draft, all authors; writing – review & editing, P.B., A.v.d.S., M.T., L.L.L., and J.P.T.-K.; visualization, P.B. and J.P.T.-K.; supervision, J.P.T.-K. and Y.B.; project administration, J.B.R.H.; funding acquisition, C.E.S.R.

### DECLARATION OF INTERESTS

All authors affiliated with Relation Therapeutics receive equity-based compensation in the company. N.V. holds stock in Glyde Bio, Inc., and Innovac Therapeutics, Inc.

Received: November 3, 2022

Revised: August 30, 2023

Accepted: September 6, 2023

Published: October 4, 2023

### REFERENCES

1. Tyers, M., and Wright, G.D. (2019). Drug combinations: a strategy to extend the life of antibiotics in the 21st century. *Nat. Rev. Microbiol.* 17, 141–155.

2. Mokhtari, R.B., Homayouni, T.S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget* 8, 38022–38043.
3. Delou, J., Souza, A.S., Souza, L., and Borges, H.L. (2019). Highlights in resistance mechanism pathways for combination therapy. *Cells* 8, 1013.
4. Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 30, 679–692.
5. Janes, J., Young, M.E., Chen, E., Rogers, N.H., Burgstaller-Muehlbacher, S., Hughes, L.D., Love, M.S., Hull, M.V., Kuhen, K.L., Woods, A.K., et al. (2018). The reframe library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proc. Natl. Acad. Sci. USA* 115, 10750–10755.
6. Clare, R.H., Bardelle, C., Harper, P., Hong, W.D., Börjesson, U., Johnston, K.L., Collier, M., Myhill, L., Cassidy, A., Plant, D., et al. (2019). Industrial scale high-throughput screening delivers multiple fast acting macrofilaricides. *Nat. Commun.* 10, 11–18.
7. Wang, H.-M., Zhang, C.-Y., Peng, K.-C., Chen, Z.-X., Su, J.-W., Li, Y.-F., Li, W.-F., Gao, Q.-Y., Zhang, S.-L., Chen, Y.-Q., et al. (2023). Using patient-derived organoids to predict locally advanced or metastatic lung cancer tumor response: A real-world study. *Cell Reports Medicine* 4, 100911.
8. Wensink, G.E., Elias, S.G., Mullenders, J., Koopman, M., Boj, S.F., Kranenburg, O.W., and Roodhart, J.M.L. (2021). Patient-derived organoids as a predictive biomarker for treatment response in cancer patients. *npj Precis. Oncol.* 5, 30.
9. Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020). Artificial Intelligence in Covid-19 Drug Repurposing (The Lancet Digital Health).
10. Sverchkov, Y., and Craven, M. (2017). A review of active learning approaches to experimental design for uncovering biological networks. *PLoS Comput. Biol.* 13, e1005466.
11. Bulusu, K.C., Guha, R., Mason, D.J., Lewis, R.P., Muratov, E., Kalantar Motamedi, Y., Cokol, M., and Bender, A. (2016). Modelling of compound combination effects and applications to efficacy and toxicity: State-of-the-art, challenges and perspectives. *Drug Discov. Today* 21, 225–238.
12. Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71, 58–63.
13. Karczewski, K.J., and Snyder, M.P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19, 299–310.
14. Wildenhain, J., Spitzer, M., Dolma, S., Jarvik, N., White, R., Roy, M., Griffiths, E., Bellows, D.S., Wright, G.D., and Tyers, M. (2015). Prediction of synergism from chemical-genetic interactions by machine learning. *Cell Syst.* 1, 383–395.
15. Cheng, F., Kovács, I.A., and Barabási, A.-L. (2019). Network-based prediction of drug combinations. *Nat. Commun.* 10, 1–11.
16. Menden, M.P., Wang, D., Mason, M.J., Szalai, B., Bulusu, K.C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., et al. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.* 10, 2674.
17. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.
18. Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, i457–i466.
19. Deac, A., Huang, Y.-H., Veličković, P., Liò, P., and Tang, J. (2019). Drug-drug adverse effect prediction with graph co-attention. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1905.00534>.
20. Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* 34, 1538–1546.
21. Jin, W., Stokes, J.M., Eastman, R.T., Itkin, Z., Zakharov, A.V., Collins, J.J., Jaakkola, T.S., and Barzilay, R. (2021). Deep learning identifies synergistic drug combinations for treating covid-19. *Proc. Natl. Acad. Sci. USA* 118, e2105070118.
22. Rozemberczki, B., Gogleva, A., Nilsson, S., Edwards, G., Nikolov, A., and Papa, E. (2021). Moomin: Deep molecular omics network for anti-cancer drug combination therapy. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2110.15087>.
23. Kashif, M., Andersson, C., Hassan, S., Karlsson, H., Senkowski, W., Fryknäs, M., Nygren, P., Larsson, R., and Gustafsson, M.G. (2015). In vitro discovery of promising anti-cancer drug combinations using iterative maximisation of a therapeutic index. *Sci. Rep.* 5, 14118.
24. Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., Malyutina, A., Jafari, M., Tanoli, Z., Pessia, A., and Tang, J. (2019). Drug-comb: an integrative cancer drug combination data portal. *Nucleic Acids Res.* 47, W43–W51.
25. Žilinskas, A., and Mockus, J. (1972). On a Bayes Method for Seeking an Extremum (Automatika i vychislitel'naja tehnika).
26. Murumägi, A., Ungureanu, D., Arjama, M., Bützow, R., Lohi, J., Sariola, H., Kanerva, J., Koskenvuo, M., and Kallioniemi, O. (2021). Strn-alk rearranged pediatric malignant peritoneal mesothelioma—functional testing of 527 cancer drugs in patient-derived cancer cells. *Transl. Oncol.* 14, 101027.
27. Murumägi, A., Ungureanu, D., Khan, S., Arjama, M., Välimäki, K., lanevski, A., lanevski, P., Bergström, R., Dini, A., Kanerva, A., et al. (2023). Drug response profiles in patient-derived cancer cells across histological subtypes of ovarian cancer: real-time therapy tailoring for a patient with low-grade serous carcinoma. *Br. J. Cancer* 128, 678–690.
28. Julkunen, H., Cichonska, A., Gautam, P., Szedmak, S., Douat, J., Pahikala, T., Aittokallio, T., and Rousu, J. (2020). Leveraging multi-way interactions for systematic prediction of pre-clinical drug combination effects. *Nat. Commun.* 11, 6136.
29. lanevski, A., Giri, A.K., Gautam, P., Kononov, A., Potdar, S., Saarela, J., Wennerberg, K., and Aittokallio, T. (2019). Prediction of drug combination effects with a minimal set of experiments. *Nat. Mach. Intell.* 1, 568–577.
30. Ling, A., and Huang, R.S. (2020). Computationally predicting clinical drug combination efficacy with cancer cell line screens and independent drug action. *Nat. Commun.* 11, 5848.
31. O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., et al. (2016). An unbiased oncology compound screen to identify novel combination strategies. *Mol. Cancer Ther.* 15, 1155–1162.
32. Holbeck, S.L., Camalier, R., Crowell, J.A., Govindharajulu, J.P., Hollingshead, M., Anderson, L.W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D., et al. (2017). The national cancer institute almanac: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* 77, 3564–3576.
33. Jaaks, P., Coker, E.A., Vis, D.J., Edwards, O., Carpenter, E.F., Leto, S.M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., et al. (2022). Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature* 603, 166–173.
34. Stumpfe, D., Hu, H., and Bajorath, J. (2020). Advances in exploring activity cliffs. *J. Comput. Aided Mol. Des.* 34, 929–942.
35. Di Veroli, G.Y., Fornari, C., Wang, D., Mollard, S., Bramhall, J.L., Richards, F.M., and Jodrell, D.I. (2016). Combenefit: an interactive platform for the analysis and visualization of drug combinations. *Bioinformatics* 32, 2866–2868.
36. Shah, H.A., Fischer, J.H., Venepalli, N.K., Danciu, O.C., Christian, S., Russell, M.J., Liu, L.C., Zacny, J.P., and Dudek, A.Z. (2019). Phase I study of aurora a kinase inhibitor alisertib (mln8237) in combination with selective vegfr inhibitor pazopanib for therapy of advanced solid tumors. *Am. J. Clin. Oncol.* 42, 413–420.
37. Serono E. Novantrone: Mitoxantrone for Injection Concentrate, Additional Safety Information. 2005. [https://www.accessdata.fda.gov/drugsatfda\\_docs/label/2009/019297s030s031lbl.pdf](https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/019297s030s031lbl.pdf).

38. Zhao, J., Quan, H., Xu, Y., Kong, X., Jin, L., and Lou, L. (2014). Flumatinib, a selective inhibitor of bcr-abl/pdgfr/kit, effectively overcomes drug resistance of certain kit mutants. *Cancer Sci.* *105*, 117–125.
39. Zhu, S., Zhang, T., Zheng, L., Liu, H., Song, W., Liu, D., Li, Z., and Pan, C.-x. (2021). Combination strategies to maximize the benefits of cancer immunotherapy. *J. Hematol. Oncol.* *14*, 156.
40. Feng, Q., Zhou, A., Zou, H., Ingle, S., May, M.T., Cai, W., Cheng, C.-Y., Yang, Z., and Tang, J. (2019). Quadruple versus triple combination antiretroviral therapies for treatment naive people with hiv: systematic review and meta-analysis of randomised controlled trials. *bmj* *366*.
41. Hirsch, C., and Schildknecht, S. (2019). In vitro research reproducibility: Keeping up high standards. *Front. Pharmacol.* *10*, 1484.
42. Avsec, Z., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Preprint at bioRxiv. <https://doi.org/10.1101/2021.04.07.438649>.
43. Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., and Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med. Genet.* *20*, 175.
44. Taylor-King, J.P. (2022). Rethinking rare disease: longevity-enhancing drug targets through x-linked aneuploidy. *Trends Genet.* *38*, 317–320.
45. Gaudalet, T., Day, B., Jamasb, A.R., Soman, J., Regep, C., Liu, G., Hayter, J.B.R., Vickers, R., Roberts, C., Tang, J., et al. (2021). Utilizing graph machine learning within drug discovery and development. *Brief. Bioinform.* *22*, bbab159.
46. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* *171*, 1437–1452.e17.
47. Chen, H., Liao, Y., Zhang, G., Sun, Z., Yang, L., Fang, X., Sun, H., Ma, L., Fu, Y., Li, J., et al. (2021). High-throughput microwell-seq 2.0 profiles massively multiplexed chemical perturbation. *Cell Discov.* *7*, 107.
48. Peidli, S., Green, T.D., Shen, C., Gross, T., Min, J., Taylor-King, J., Marks, D., Luna, A., Bluthgen, N., and Sander, C. (2022). scPerturb: Information resource for harmonized single-cell perturbation data. Preprint at bioRxiv. <https://doi.org/10.1101/2022.08.20.504663>.
49. Bray, M.-A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., and Carpenter, A.E. (2016). Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* *11*, 1757–1774.
50. Nguyen, T.-M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.* *20*, 203–215.
51. Taylor-King, J.P., Baratchart, E., Dhawan, A., Coker, E.A., Rye, I.H., Russnes, H., Chapman, S.J., Basanta, D., and Marusyk, A. (2019). Simulated ablation for detection of cells impacting paracrine signalling in histology analysis. *Math. Med. Biol.* *36*, 93–112.
52. Ribeiro, R.P., and Moniz, N. (2020). Imbalanced regression and extreme value prediction. *Mach. Learn.* *109*, 1803–1835.
53. Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. (2018). Film: Visual reasoning with a general conditioning layer. *Proc. AAAI Conf. Artif. Intell.* *32*.
54. Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems30* (Curran Associates, Inc.).
55. Jain, M., Lahlou, S., Nekoei, H., Butoi, V., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. (2021). DEUP: Direct Epistemic Uncertainty Prediction (CoRR). [abs/2102.08501](https://arxiv.org/abs/2102.08501).
56. Borkowski, O., Koch, M., Zettor, A., Pandi, A., Batista, A.C., Soudier, P., and Faulon, J.-L. (2020). Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat. Commun.* *11*, 1872.
57. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* *427*, 247–252.
58. Carbonell, P., Jervis, A.J., Robinson, C.J., Yan, C., Dunstan, M., Swainston, N., Vinaixa, M., Hollywood, K.A., Currin, A., Rattray, N.J.W., et al. (2018). An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol.* *1*, 66.
59. Hie, B., Bryson, B.D., and Berger, B. (2020). Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Syst.* *11*, 461–477.e9.
60. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., and Overington, J.P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* *43*, W612–W620.
61. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2019). Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.* *47*, D1102–D1109.
62. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* *569*, 503–508.
63. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* *12*, 2825–2830.
64. RDKit: Open-Source Cheminformatics. <https://doi.org/10.5281/zenodo.7357998>
65. Morgan, H.L. (1965). The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J. Chem. Doc.* *5*, 107–113.

Resource

# Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction

Markus Müller,<sup>1,2,3,4,\*</sup> Florian Huber,<sup>1,2,3</sup> Marion Arnaud,<sup>1,2,3</sup> Anne I. Kraemer,<sup>1,2,3</sup> Emma Ricart Altimiras,<sup>1,2,3</sup> Justine Michaux,<sup>1,2,3</sup> Marie Taillandier-Coindard,<sup>1,2,3</sup> Johanna Chiffelle,<sup>1,2,3</sup> Baptiste Murgues,<sup>1,2,3</sup> Talita Gehret,<sup>1,2,3</sup> Aymeric Auger,<sup>1,2,3</sup> Brian J. Stevenson,<sup>3,4</sup> George Coukos,<sup>1,2,3,5</sup> Alexandre Harari,<sup>1,2,3,5</sup> and Michal Bassani-Sternberg<sup>1,2,3,5,6,\*</sup>

<sup>1</sup>Ludwig Institute for Cancer Research, University of Lausanne, Agora Center Bugnon 25A, 1005 Lausanne, Switzerland

<sup>2</sup>Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Rue du Bugnon 46, 1005 Lausanne, Switzerland

<sup>3</sup>Agora Cancer Research Centre, 1011 Lausanne, Switzerland

<sup>4</sup>SIB Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland

<sup>5</sup>Center of Experimental Therapeutics, Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Rue du Bugnon 46, 1005 Lausanne, Switzerland

<sup>6</sup>Lead contact

\*Correspondence: [markus.muller@chuv.ch](mailto:markus.muller@chuv.ch) (M.M.), [michal.bassani@chuv.ch](mailto:michal.bassani@chuv.ch) (M.B.-S.)

<https://doi.org/10.1016/j.immuni.2023.09.002>

## SUMMARY

The accurate selection of neoantigens that bind to class I human leukocyte antigen (HLA) and are recognized by autologous T cells is a crucial step in many cancer immunotherapy pipelines. We reprocessed whole-exome sequencing and RNA sequencing (RNA-seq) data from 120 cancer patients from two external large-scale neoantigen immunogenicity screening assays combined with an in-house dataset of 11 patients and identified 46,017 somatic single-nucleotide variant mutations and 1,781,445 neo-peptides, of which 212 mutations and 178 neo-peptides were immunogenic. Beyond features commonly used for neoantigen prioritization, factors such as the location of neo-peptides within protein HLA presentation hotspots, binding promiscuity, and the role of the mutated gene in oncogenicity were predictive for immunogenicity. The classifiers accurately predicted neoantigen immunogenicity across datasets and improved their ranking by up to 30%. Besides insights into machine learning methods for neoantigen ranking, we have provided homogenized datasets valuable for developing and benchmarking companion algorithms for neoantigen-based immunotherapies.

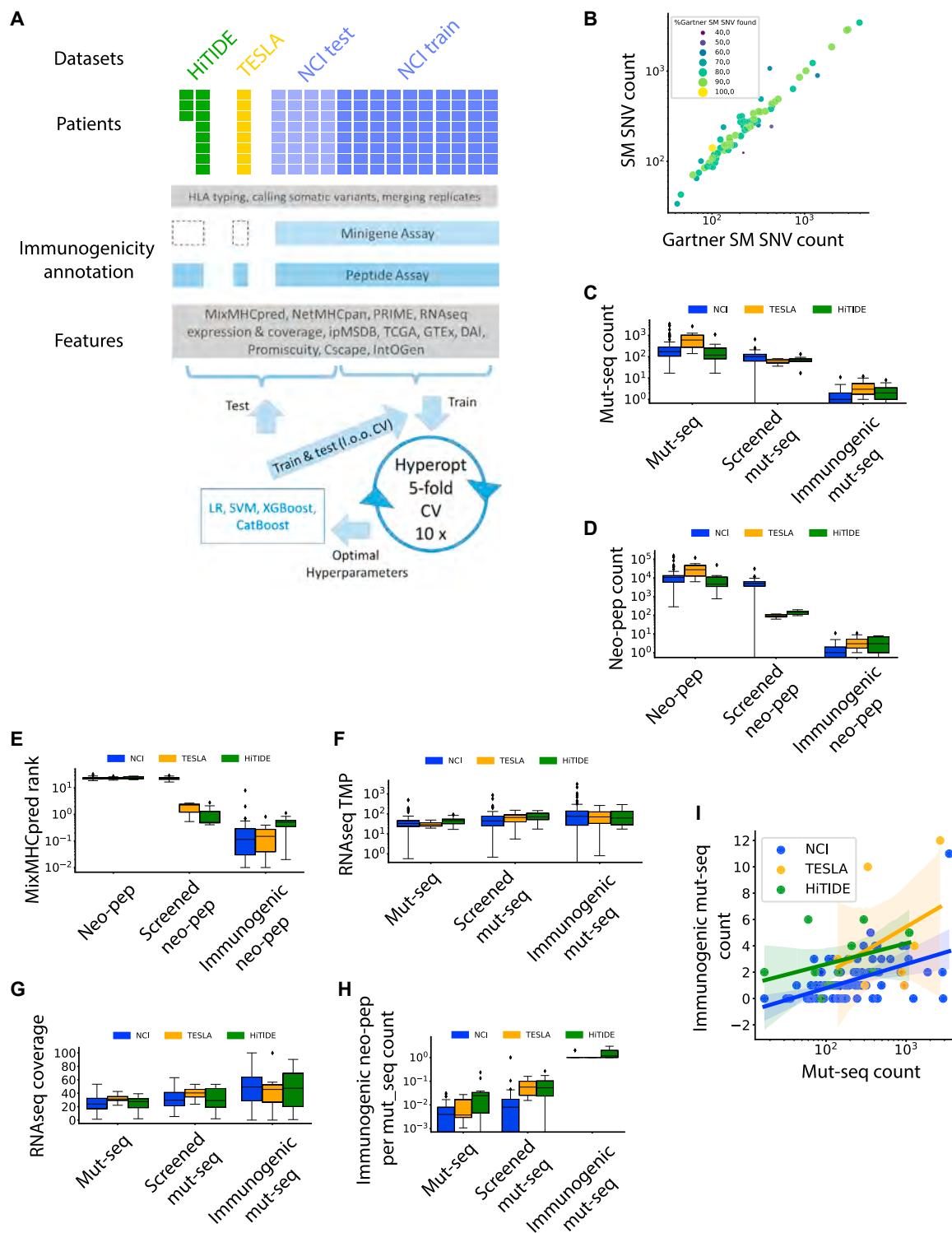
## INTRODUCTION

In recent years, it has been demonstrated across tumor types in patients receiving adoptive transfer of autologous *in vitro* cultured tumor infiltrating lymphocytes (TILs) that T cells specifically recognizing mutated neoantigens play a key role in mediating effective anti-tumor responses.<sup>1–3</sup> Furthermore, neoantigens are found to be implicated in the therapeutic efficacy of immune checkpoint inhibitor antibodies,<sup>4,5</sup> and several studies show immune recognition following neoantigen-based vaccines,<sup>6,7</sup> where patients experience no major toxicities.

Mutated proteins are processed and presented on tumor cells as human leukocyte antigen (HLA) binding peptides (HLAp) and are recognized by cognate T cell receptors (TCRs) as “non-self.” Targeting such neoantigens enables immune cells to distinguish between normal and cancerous cells, diminishing the risk of autoimmunity. Technological improvements in genomics, bioinformatics, and *in silico* HLA binding prediction tools have facilitated breakthroughs in the discovery of neoantigens encoded by somatic non-synonymous single-nucleotide variants (SNVs), insertions and deletions (InDels), and frameshifts (FSs) that arise

during the process of tumorigenesis and are not expressed by normal cells.<sup>8,9</sup> Furthermore, advanced immunological screening techniques have facilitated the detection and isolation of neoantigen reactive T cells.<sup>10–13</sup>

The development of innovative clinical treatment options targeting neoantigens requires the identification of neoantigens that are targeted by autologous T cells. However, only a small percentage of neoantigens are immunogenic, which makes their identification challenging.<sup>14</sup> Various algorithms that score and rank neoantigens based on their likelihood of being presented on the patient’s HLA molecule<sup>15–18</sup> and being specifically recognized by high avidity T cell clonotypes<sup>19–23</sup> have been proposed. Other groups have provided pipelines for mutation detection and neoantigen prioritization.<sup>24,25</sup> Despite all these efforts, a recent study shows little consensus in the neoantigen ranking performed by different laboratories,<sup>26</sup> and the performance of immunogenicity prediction methods varies between different datasets.<sup>14</sup> As datasets with hundreds or thousands of neoantigen immunogenicity measurements become available,<sup>26–28</sup> machine learning (ML) methods are able to train powerful immunogenicity prediction algorithms taking into account the multidimensional



**Figure 1. Statistics reveal the reproducibility of our pipeline and the bias in mutation and neo-peptide subsets**

(A) Data processing workflow applied in this paper. WES and RNA-seq data were downloaded and processed. Mutations and neo-peptides were annotated with the results from the immunogenicity screens, and the feature scores and annotations were added. The NCI data matrix was split into train- and test sets, and the classifiers were trained on the subset of screened mutations or neo-peptides in NCI-train (see STAR Methods for naming rules for the data subsets) using Hyperopt parameter optimization and 5-fold cross validation (CV) in 10 replicate runs. The trained classifiers were tested on all neo-peptides or mutations in NCI-train (using leave one out CV), NCI-test, TESLA, and HITIDE cohorts.

(legend continued on next page)



structure of the data. In a recent example, the ranking based on an ML model has outperformed a ranking based on binding affinity only.<sup>28</sup> This improvement in prioritizing immunogenic neoantigens is particularly important for neoantigen or mRNA vaccines, where only a limited set of neoantigens are included.<sup>2,3,6,7</sup>

Here, we studied the performance of state-of-the-art ML algorithms using two public datasets (National Cancer Institute [NCI] with 112 patients<sup>27,28</sup> and Tumor Neoantigen Selection Alliance [TESLA] with 8 patients<sup>26</sup>) plus an additional in-house dataset composed of 11 patients, 2 of which were already included in a previous publication.<sup>13</sup> We reprocessed all whole-exome sequencing (WES) and RNA sequencing (RNA-seq) data with a uniform mutation detection pipeline and investigated the robustness of different ML algorithms and data preprocessing steps. We demonstrated that classifiers trained on the large NCI dataset can accurately predict the immunogenicity of neoantigens on each test dataset. With orthogonal features, our ML based approach outperformed previously published methods<sup>28</sup> and increased the number of immunogenic peptides ranked in the top 20 by 30%. Compared with the ranking reported in the TESLA consortium study,<sup>26</sup> our ML methods performed favorably and came first in two out of three ranking evaluation metrics. We provide classifiers and data processing methods for the improved prioritization of immunogenic neoantigens. The uniformly processed datasets are unique resources for other groups active in the field of immunogenicity prediction and in the development of innovative neoantigen-based therapies.

## RESULTS

### Our mutation detection is consistent with published results

Cancer cells can have several hundred somatic mutations (SMs), but only a few of them may be presented as HLA binding neo-peptides and recognized by T cells. The accurate selection of a limited number of mutations (e.g., for mRNA cancer vaccine) or neo-peptides (e.g., for multimer based sorting of neoantigen-specific T cells) that are most likely to be immunogenic is a crucial step in cancer vaccines and adoptive transfer of T cells. Here, we used two public (NCI<sup>27,28</sup> and TESLA<sup>26</sup>) and one in-house (the Human Integrated Tumor Immunology Discovery Engine; HiTIDE) dataset to train and test ML algorithms for the prioritization of mutations and neo-peptides (Figure 1A). The datasets consisted of WES and RNA-seq data as well as immunogenicity assay results for hundreds of mutations and/or neo-

peptides (Table 1; Data S1). The main difference between the datasets laid in the way mutations (*mut-seq*, typically 25 amino acid (aa) sequences with mutation in the center) or neo-peptides (*neo-pep*, peptides of length 8–12 including mutation) were selected for immunogenicity screening and in the screening methods used (STAR Methods). In the NCI dataset, many mutations and neo-peptides were physically screened as reported by Gartner et al.<sup>28</sup> In a cohort of 112 patients, which we defined here as *NCI\_mut-seq*, for almost all the expressed mutations, minigenes encoding the mutations and 12 flanking wild-type (WT) aa on each side were transcribed *in vitro* and transfected into autologous antigens presenting cells (APCs) followed by a co-culture with TIL cultures and interferon (IFN)- $\gamma$  enzyme-linked immunospot (ELISpot) immunogenicity measurement. For 80 of the 112 patients, a cohort we defined as the *NCI\_neo-pep*, additional immunogenicity screens were performed to identify the optimal neo-antigenic epitopes and their HLA restrictions. The top-ranked neo-peptides predicted by NetMHCpan to span immunogenic mutations from the above mini-gene assay were pulsed on autologous APCs or APCs engineered to express the patient's HLA-I alleles, prior to co-culture with TILs and IFN- $\gamma$  ELISpot readout. Neo-peptides with positive ELISpot readout were assigned as immunogenic. All other neo-peptides containing the immunogenic mutation and all neo-peptides containing screened non-immunogenic mutations were considered as non-immunogenic. In the TESLA study, immunogenicity of selected neo-peptides was determined with labeling of subject-matched TILs or peripheral blood mononuclear cells (PBMCs) with HLA-I peptide multimers.<sup>26</sup> The immunogenicity of selected neo-peptides in the HiTIDE cohort was interrogated with IFN- $\gamma$  ELISpot assays following incubation of the peptides with either bulk TILs or neo-antigen enriched TILs (NeoScreen method) grown from tumor biopsies in the presence of APCs loaded with neo-peptides (Figures S1A and S1B), as previously described.<sup>13</sup> Importantly, in the TESLA and HiTIDE datasets, only a selection of neo-peptides was experimentally screened, and the immunogenicity annotation of the mutations was inferred accordingly.

First, we uniformly processed all data, conducting HLA typing, mutation calling, RNA-seq gene expression analysis, and read coverage assessment at the specific loci of the SM. To assure capturing all relevant mutations in the NCI dataset, prior to the ML training, we assessed the extent to which we were able to reproduce the genomic analysis results published by Gartner et al.<sup>28</sup> First, for a subset of 80 of the 112 patients, for which HLA typing results from Gartner et al. were available, we

(B) Comparison of SM SNV mutation counts obtained from Gartner et al.<sup>28</sup> and our analysis for a subset of 80 patients, where each patient corresponds to a data point. The size and color of the points reflect the percentage of mutations identified by Gartner et al. that were also identified in our analysis. (C) to (H) show different statistics of the patient data for the NCI, TESLA, and HiTIDE datasets, where only SM SNVs were considered. The statistics were obtained from all mutations or neo-peptides per patient (left group of boxes), or from the subsets of screened and immunogenic mutations or neo-peptides per patient (middle and right group of boxes).

(C) Mutation counts.

(D) Neo-peptide counts. The outliers in (C) and (D) originate from NCI patients for which all non-immunogenic peptides were annotated as not-screened in Gartner et al.<sup>28</sup>

(E) Average MixMHCpred %rank scores.

(F) Average RNA expression in TPM.

(G) Average RNA coverage of the mutations in %.

(H) Average number of immunogenic neo-peptides per mutation.

(I) Immunogenic mutation counts as a function of the mutation counts for each patient in the NCI, TESLA, and HiTIDE datasets.

See also Figure S1.

**Table 1. Mutation and neo-peptide counts for the NCI, TESLA, and HiTIDE datasets and their subsets**

	Number of patients (train or test datasets)	Immunogenicity screening method	Immunogenic	Not immunogenic	Not screened
NCI mutations	89 (train), 23 (test)	minigenes, IFNg ELISpot	146	11,651	24,899
NCI neo-peptides	57 (train), 23 (test)	peptides, IFNg ELISpot	103	418,872 <sup>a</sup>	953,486
TESLA mutations	8 (test)	<i>in silico</i>	36 <sup>b,c</sup>	461 <sup>b</sup>	6,231 <sup>b</sup>
TESLA neo-peptides	8 (test)	peptides, HLA-I peptide multimers	34 <sup>c</sup>	702	300,505
HiTIDE mutations	11 (test)	<i>in silico</i>	30 <sup>b</sup>	751 <sup>b</sup>	1,812 <sup>b</sup>
HiTIDE neo-peptides	11 (test)	peptides, IFNg ELISpot	41	1,511	106,191

<sup>a</sup>Non-immunogenic NCI neo-peptide dataset contains three types of neo-peptides: the screened ones with negative immunogenicity test, the inferred ones from immunogenic mutations that were not screened at the neo-peptide level, and the ones from screened but non-immunogenic mutations.

<sup>b</sup>Immunogenicity of mutations in TESLA and HiTIDE datasets was inferred from the immunogenicity screens of the respective neo-peptides.

<sup>c</sup>Neo-peptides are excluded if they match WT peptide or missing value cannot be imputed resulting in more mutations than neo-peptides.

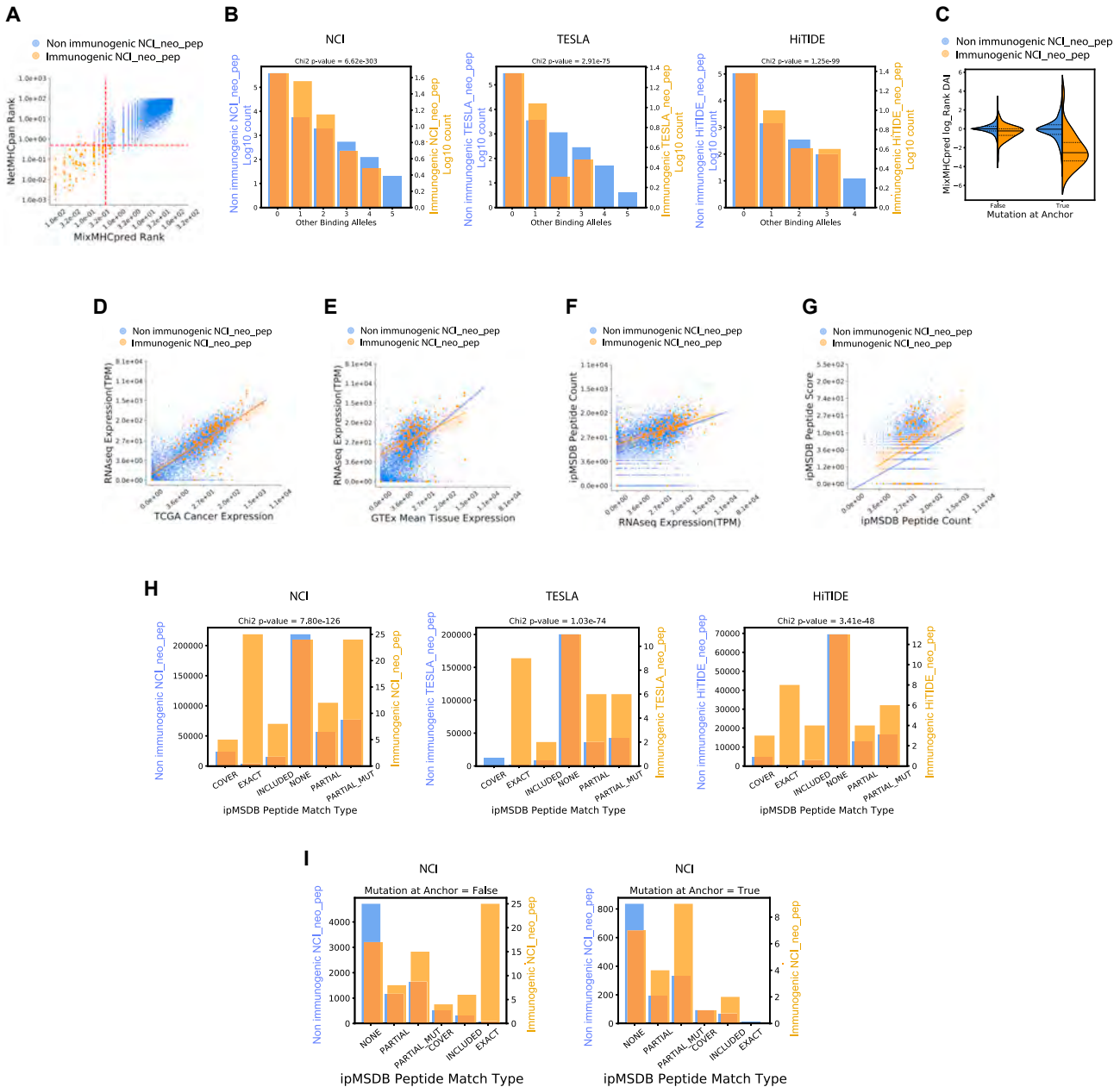
compared the HLA allotypes. The HLA typing was overall consistent, and we found that, for 74 patients, all alleles were identical, for 4 patients, 1 or 2 alleles were missed by us or by Gartner et al., and in 2 patients, there were conflicting alleles but with similar sequence motifs (Data S2). In addition, the SNV SM counts we obtained correlated well with the counts reported by Gartner et al. (Figure 1B; Data S2). Overall, in the subset of 80 patients we identified 31,880 SNV SMs, including 82.2% (26,420 out of 32,148) of the SNV SMs published by Gartner et al., where 67.5% of the patients (54 out of 80) had a SNV SM overlap larger than 80% (Data S2). For a few patients there was a substantial variation in the number of mutations detected and for two patients we called less than 50% of the mutations reported by Gartner et al. Interestingly, we detected 143 of the 151 (94.7%) immunogenic SNV SMs published in Gartner et al. (Data S2; Figure S1C). Good correlations were also obtained when we compared insertion and deletion mutations with and without FSs (Figures S1D and S1E).

### Immunogenicity-related feature scores highlight subtle differences between datasets

Next, we added multiple feature scores (Data S3) reflecting the propensity of a peptide to be presented, such as bulk RNA-seq gene expression of the mutated gene and its expression in the tissue-matched Cancer Genome Atlas (TCGA) (<https://www.cancer.gov/tcga>) and the tissue-matched Genotype-Tissue Expression (GTEx) atlas (<https://gtexportal.org/>), proteasomal cleavage scores,<sup>29</sup> tapasin binding,<sup>30</sup> binding affinity to HLA-I allotypes (NetMHCpan,<sup>15</sup> MixMHCpred<sup>16</sup>), and stability ranks.<sup>31</sup> Other feature scores evaluated the dissimilarity of a neo-peptide to the WT peptide counterpart (differential agretopicity index or DAI)<sup>19,22,32,33</sup> and the potential of a neoantigen to bind several alleles. A notable bias toward hydrophobic aa was observed at T cell receptor contact residues within immunogenic epitopes.<sup>34</sup> We therefore employed also the PRIME predictor, that captures such hydrophobicity related molecular properties associated with TCR recognition.<sup>23,35</sup> We also used our large-scale in-house immunopeptidome database (ipMSDB<sup>36</sup>) of HLA-bound WT peptides identified by mass spectrometry (MS) to assess the likelihood of neo-peptides to be naturally processed and presented at the cell surface by HLA (see below and in STAR Methods).

Last, it is well established that mutations in oncogenes and tumor suppressors are enriched across cancers and specific sites are more frequently mutated. Hoyos et al. has modeled the relationship between oncogenicity and immunogenicity for tumor driver mutations, focusing on p53 mutations, and demonstrated that hotspot mutations optimally solve an evolutionary trade-off between oncogenic potential and neoantigen immunogenicity.<sup>37</sup> Therefore, we scored SNV SM based on their appearance in the population with the Integrative Onco Genomics (IntOGen) database,<sup>38</sup> and we predicted their oncogenic status (disease-driver or neutral) with the CScape tool<sup>39</sup> to assess the role of the mutation in tumorigenesis.

Comparison of basic statistics across all three datasets (Data S1) revealed that the number of SNV SMs called per patient was highest for the TESLA dataset (Figure 1C), which contained only melanoma and non-small cell lung cancer (NSCLC) samples that are known for high mutational loads. In contrast, the number of mutations per patient screened with the mini-gene approach in the NCI dataset was higher than the mutations included in neo-peptide screens in the TESLA and HiTIDE datasets. The number of immunogenic mutations per patient was higher in TESLA and HiTIDE, possibly because of differences in cancer types and the sensitivity of immunogenicity screening methods. In the NCI dataset—following the annotations provided in Gartner et al.<sup>28</sup>—all neo-peptides originating from screened mutations were considered as screened, even if only the mutation, but not the neo-peptide was actually screened. Therefore, the number of neo-peptides annotated as “screened” was much higher in the NCI dataset (Figure 1D), and there was no difference in binding affinity between screened and not-screened neo-peptides (Figure 1E). In contrast, binding affinity was used as a screening criterion in the TESLA and HiTIDE datasets (Figure 1E). The RNA-seq gene expression values revealed small differences between datasets. In all datasets, mutations selected for T cell screening had higher RNA-seq gene expression, and this effect was strongest in the HiTIDE- and weakest in NCI data (Figure 1F). RNA-seq mutation coverage was consistently employed as a screening criterion in all datasets, with the TESLA dataset demonstrating the most pronounced utilization of this filter (Figure 1G). The number of immunogenic neo-peptides per mutation was higher in HiTIDE and TESLA datasets (Figure 1H). In the NCI and TESLA datasets, on average only one immunogenic



**Figure 2. Exploring relationships between features and predictive value for immunogenicity**

Scatterplots display the immunogenic (orange) and non-immunogenic (blue) neo-peptides or mutations with their regression lines for the screened *NCI\_neo-pep/mut-seq* dataset. Only a random subsample of 10,000 points of the non-immunogenic points is shown in the scatterplots. Histograms display the feature scores of immunogenic (orange) and non-immunogenic (blue) neo-peptides for the screened *NCI\_neo-pep*, *TESLA\_neo-pep*, and *HITIDE\_neo-pep* datasets. The scale of the immunogenic neo-peptide counts is given on the right y axis; the scale of the non-immunogenic counts is on the left y axis. The p values shown in the histogram titles evaluate the difference between immunogenic and non-immunogenic feature values and are calculated by a  $\chi^2$  test.

- (A) Scatterplot of MixMHCpred and NetMHCpan %rank scores. Red dashed lines mark the 0.5% ranks.  
 (B) Histogram for "Number Binding Alleles" scores. Note the different log-scales for immunogenic and non-immunogenic neo-peptides counts.  
 (C) Violin plot of MixMHCpred log-rank DAI for neo-peptides with mutations at anchor and non-anchor positions.  
 (D) TCGA expression versus RNA-seq expression.  
 (E) GTEx expression versus RNA-seq expression.  
 (F) Scatterplot of *ipMSDB Peptide Count* per protein versus RNA-seq expression.  
 (G) *ipMSDB Peptide Count* per protein versus *ipMSDB Peptide Score*.

(legend continued on next page)

neo-peptide was detected per immunogenic mutation, whereas in the HiTIDE cohort this number was slightly higher. The number of immunogenic neo-peptides per patient correlated with the total number of SNV SMs detected in a patient (Figure 1I). In summary, the NCI dataset had the highest number of screened mutations and neo-peptides with the least selection bias and is therefore most suitable for training ML models.

### Features beyond binding affinity and gene expression correlate with immunogenicity

Next, we investigated how the mutation or neo-peptide features correlated with immunogenicity. By examining these correlations, we sought to gain insights into the factors that contribute to immunogenicity and potentially identify key determinants of immune recognition. We found that, in agreement with published results,<sup>26,28,40</sup> features describing proteasomal cleavage, transporter associated with antigen presentation (TAP) import into endoplasmic reticulum and binding stability, correlated with immunogenicity in all three datasets, and they correlated poorly with binding affinity (Figures S2A–S2C). In addition, as previously demonstrated,<sup>26,28</sup> features reflecting the binding affinity between a neo-peptide and the patients' HLA-I alleles were among the strongest predictors for immunogenicity for all three datasets (Figure S2D). Although NetMHCpan and MixMHCpred prediction %rank scores correlated, they contained complementary information. For example, in the NCI dataset, ten immunogenic neo-peptides did not pass the binding threshold of %rank  $\leq 0.5$  with NetMHCpan, but they passed it with MixMHCpred (Figure 2A). We found that promiscuous neo-peptides that were predicted to bind to multiple patient's HLA-I alleles were more likely to be immunogenic than neo-peptides predicted to bind a single allele (Figure 2B), possibly because binding to multiple alleles increases the chance for HLA-I presentation and makes the presentation of neo-peptides more resistant to loss of specific HLA-I alleles.<sup>41</sup> Along the same lines, mutations with a higher number of neo-peptides weakly binding to a patient's HLA-I alleles, were more likely to be immunogenic (Figure S2E). The PRIME prediction rank differences between immunogenic and non-immunogenic neo-peptides were similar to those of MixMHCpred or NetMHCpan (Figure S2F). DAI values for binding prediction log-ranks were lower for immunogenic neo-peptides (Figure S2G) in agreement with previous results.<sup>19,22,33</sup> As expected, the location of mutations in an anchor position was not significant per se (Figure S2H), but it became important in combination with DAI values, which were significantly lower (t test p value  $2.19 \times 10^{-17}$ ) for immunogenic mutations at anchor positions (Figure 2C). Based on the analyzed data, there was no obvious tendency for mutations to be placed in the middle of a neo-peptide, and the enrichment of immunogenic mutations in the middle of 10 mers reported for the TESLA dataset<sup>26</sup> could not be confirmed for the NCI and HiTIDE datasets (Figure S2I). As expected, immunogenic neo-peptides were strongly enriched in the group of 9 or 10 mer peptides, reflecting the length preferences of HLA-I alleles (Figure S2J). HLA binding-aff-

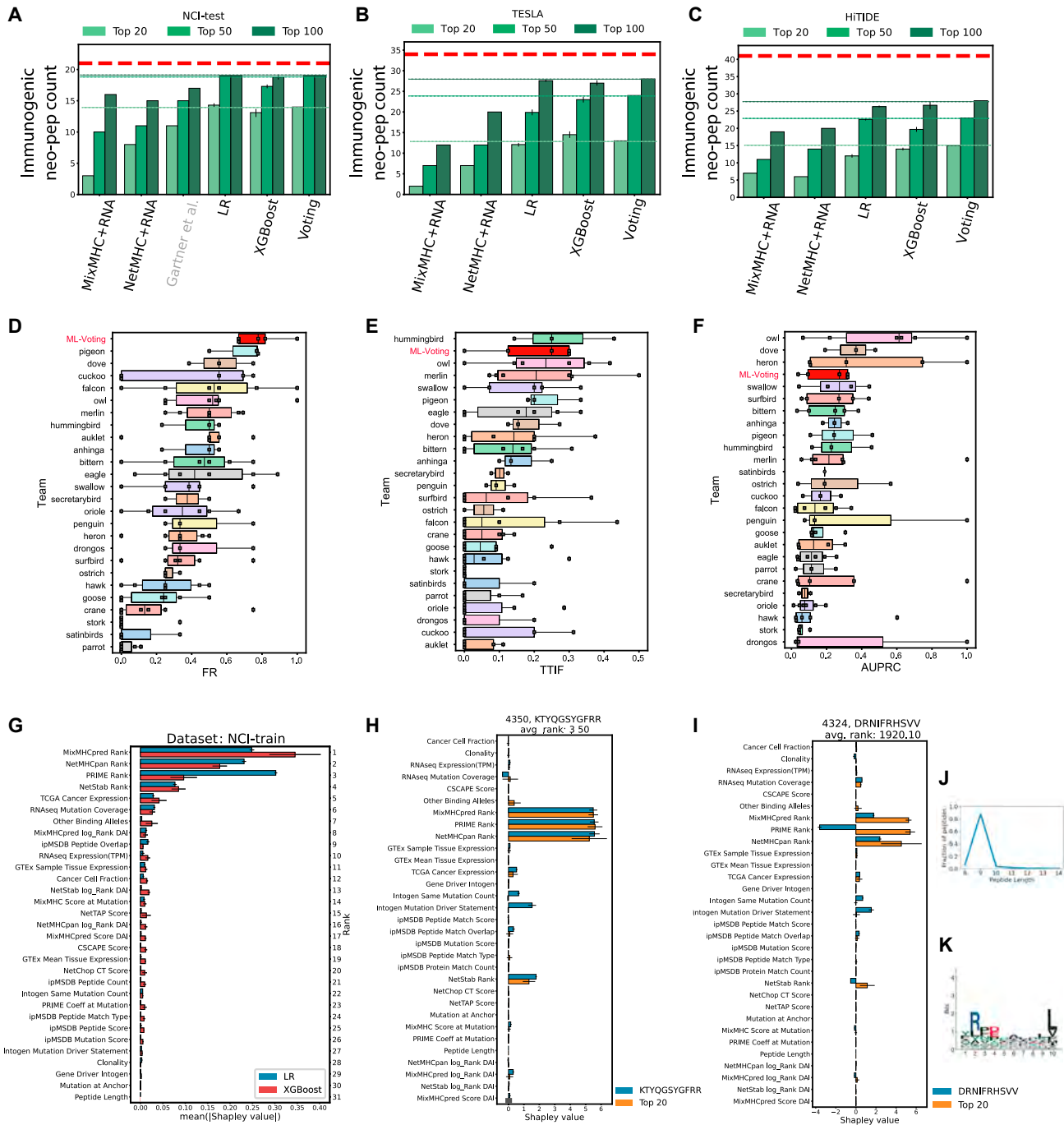
inity predictors that incorporate peptide length preferences were used to select the neo-peptides for immunogenicity testing. Hence, based on these three datasets, it is challenging to determine whether this enrichment stems from a bias in selecting neo-peptides or if it represents an intrinsic characteristic of immunogenic peptides. It has been demonstrated that gene or protein expression positively impacts HLA-I presentation<sup>40,42</sup> and immunogenicity.<sup>26,28</sup> In all three datasets, immunogenic mutations had higher gene expression and higher coverage of the mutation in the patient's tumor bulk RNA-seq data compared with non-immunogenic ones (Figures S3A and S3B). To investigate the possibility of substituting a patient's gene expression values with data from publicly available datasets, particularly in scenarios where the patient's tumor tissue RNA-seq data are unavailable, we included tissue-matched RNA-seq expression data from the TCGA and GTEx repositories as additional features. For both immunogenic and non-immunogenic mutations, the TCGA gene expression correlated strongly (Pearson's  $R = 0.818$ ) with its expression in the patient's cancer tissue (Figure 2D). The gene expression in GTEx correlated to a lower extent (Pearson's  $R = 0.645$ ), and the regression line for immunogenic mutations was shifted to higher RNA-seq values compared with the regression line for non-immunogenic ones (Figure 2E). We concluded that immunogenic mutated genes had higher gene expression in cancer tissues compared with the matched healthy tissues in GTEx, and the expression values were better captured by TCGA. Lastly, cancer cell fraction (CCF), clonality, and zygosity were not associated with immunogenicity (Data S3).

Our in-house ipMSDB database<sup>36</sup> contains WT HLA-I and -II ligands identified by MS in multiple healthy and cancerous human tissues and cell lines with various HLA allotypes. The ipMSDB version used in this work contains 547,476 unique HLA-I peptides, which we used to infer the HLA-I presentation of a neoantigen based on the coverage of the corresponding WT peptide and on the natural presentation of the source protein. We found that the number of ipMSDB peptides mapped to a protein ("*ipMSDB Peptide Count*") was significantly higher for proteins containing immunogenic mutations across all three datasets (Figure S3C). These data indicate that immunogenic peptides in the three datasets preferably belong to proteins that are naturally processed and presented, in agreement with previous findings.<sup>18,36,43</sup> *ipMSDB Peptide Count* for a given protein correlated (Pearson's  $R = 0.498$ ) with mRNA expression of the corresponding gene (Figure 2F), but this correlation could not fully explain the higher *ipMSDB Peptide Count* values for immunogenic mutations (Figure S3D), suggesting that these features are not fully redundant. In addition, the "*ipMSDB Peptide Score*" measures the overlap between the WT peptide within ipMSDB and the neo-peptides (Figure S3E). The correlation between the *ipMSDB Peptide Score* and the *ipMSDB Peptide Count* (Pearson's  $R = 0.435$ ) reflects that proteins with overall more ipMSDB peptides had a better chance to cover a neo-peptide. However, the *ipMSDB Peptide Score* was higher for

(H) Histograms for "*ipMSDB Peptide Match Type*."

(I) Histograms for *ipMSDB Peptide Match Type* for neo-peptides with or without a mutation at an anchor position of the HLA allele with the lowest MixMHCpred % rank score.

See also Figures S2 and S3.



**Figure 3. Assessments of the classifier's performance and feature importance**

(A) Immunogenic neo-peptides were ranked per patient and the number of immunogenic neo-peptides in the top 20, 50, or 100 ranks was calculated and summed up for all patients in the NCI-test dataset. The ranking was performed either by NetMHCpan and RNA expression, MixMHCpred and RNA expression as described in the text, or logistic regression (LR), XGBoost, or the voting classifier. "Gartner et al." refers to the ranking reported in Gartner et al.<sup>28</sup> The red dashed horizontal lines indicate the total number of immunogenic neo-peptides in NCI-test. The green lines mark the median performance of the voting classifier in the top 20, 50, or 100 ranks according to their respective colors.

(B) As in (A), but for the TESLA dataset.

(C) As in (A), but for the HITIDE dataset.

(D) Comparison of the fraction ranked (FR) score obtained by the voting classifier trained on NCI-train and tested on TESLA. FR scores of the TESLA participants were obtained from Wells et al.<sup>26</sup> The FR score gives us the fraction of immunogenic neo-peptides ranked in the top 100 per patient.

(E) Same as (D) but for the top-20 immunogenic fraction (TTIF) score. The TTIF score gives us the fraction of immunogenic neo-peptides among all screened neo-peptides ranked in the top 20 per patient.

(legend continued on next page)

immunogenic neo-peptides compared with non-immunogenic ones (Figure 2G), and this shift was significant in all three datasets (Figure S3F). We also found a highly significant enrichment of immunogenic neo-peptides, which either mapped exactly to the WT counterpart sequences in ipMSDB or were fully included in such sequences (Figure 2H). These results indicated that immunogenic neo-peptides were preferably found in HLA-I presentation “hotspots” and that utilizing sequence matching to ipMSDB proves to be an effective strategy for prioritizing “true” HLA-I binding neo-peptides, as long as the mutation does not occur in an anchor position (Figure 2I). When mutations arise in anchor positions, they tend to produce a predicted peptide variant that exhibits superior binding affinity compared with the original WT peptide especially for immunogenic peptides (Figure 2C). Consequently, in these scenarios, the likelihood of finding the WT peptide represented in the ipMSDB is reduced (Figure 2I).

Further, we included features that evaluate the impact of a mutation on the cellular or molecular function of the mutated protein. Although Cscape<sup>39</sup> is an oncogenicity predictor, we demonstrated that it had also a predictive value for immunogenicity (Figure S3G), possibly because oncogenic mutations often destabilize the protein structure, leading to rapid degradation of the protein and presentation on HLA-I.<sup>44</sup> We also included mutation annotations from the IntOGen<sup>38</sup> database, and we further found that mutations annotated as oncogenic drivers were enriched for immunogenicity in all three datasets (Figure S3H), and there was a slight immunogenicity enrichment for mutations with a lower prevalence in the population (Figure S3I).

### Classifiers trained on a large unbiased dataset accurately rank neo-peptides in other datasets

Neoantigen-based personalized immunotherapy strategies rely on the selection of the most promising mutations or neo-peptides. For both mutations and neo-peptides, we trained a separate ML model, which calculates the probability that a mutation or neo-peptide can induce a spontaneous immune response, as was captured by the immunogenicity screening assays, and this probability is then used for the ranking. First, we investigated the ranking of neo-peptides. We used the Bayesian optimization framework Hyperopt<sup>45</sup> to train the classifiers and their hyperparameters on NCI-train (Figure 1A; see supplemental information for the details). Through leave-one-out cross-validation (CV) testing on the NCI-train dataset, we observed that the logistic regression (LR)<sup>46</sup> classifier’s performance showed improvement as the number of non-immunogenic neo-peptides in the training set increased (Figure S4A). Additionally, increasing the

number of Hyperopt iterations also contributed to the enhanced performance of the LR classifier (Figure S4B). These findings highlight the importance of a larger training set and extensive Hyperopt iterations in optimizing the performance of the LR and other classifiers for neo-peptide immunogenicity prediction.

Furthermore, the choice of data normalization method had an impact on the performance of the LR classifier, as demonstrated by Figure S4C. Notably, employing quantile normalization resulted in a remarkable 134.0% increase in the number of immunogenic neo-peptides ranked within the top 20, in comparison with the scenario where no normalization was applied (Figure S4D). These findings underscore the importance of implementing appropriate data normalization techniques, such as quantile normalization, to enhance the accuracy and predictive power of the LR classifier.

Furthermore, the choice of classifier algorithm had an impact on the number of immunogenic neo-peptides ranked among the top positions (Figure S4E). For NCI-train with leave-one-out CV, LR performed best, followed by XGBoost,<sup>47</sup> CatBoost,<sup>48</sup> and the SVMs.<sup>49</sup> The LR classifier was able to rank 49.1% of immunogenic neo-peptides in the top 20, 62.2% in the top 50, and 75.6% in the top 100 (Figure S4F; Data S4). The principal-component analysis (PCA) plot (Figure S4G) revealed that LR and XGBoost produce distinct and complementary rankings. The plot visually demonstrated that these LR and XGBoost offer diverse perspectives and capture different aspects of neo-peptide immunogenicity, indicating the potential benefit of leveraging their combined results for a more comprehensive and accurate assessment of immunogenic rankings. Therefore, we constructed a voting classifier, which averaged the immunogenic class probabilities of all ten LR and ten XGBoost classifier replicates (STAR Methods). Across the NCI-test, TESLA, and HiTIDE test datasets, the ranking of the voting classifier was always better or comparable to the rankings of the LR and XGBoost classifiers (Figures 3A–3C). We concluded that the voting classifier provides a ranking that is more robust and less dependent on the dataset.

The performance of ML ranking can vary depending on the dataset used. To investigate this further we trained and tested the LR classifier on HiTIDE with leave-one-out CV (see STAR Methods) and compared it with the LR classifier trained on the much larger NCI-train dataset. The HiTIDE-trained LR performed clearly better on HiTIDE neo-peptides, but it performed worse on the TESLA and NCI-test datasets (Figures S4H–S4J). The LR classifiers had a preference for features such as RNA-seq expression, CCF, and ipMSDB scores, which were used in the HiTIDE cohort to select neo-peptides for immunogenicity screening (Figure S4K). These findings demonstrated that ML

(F) Same as (D) but for the “area under the precision recall curve” (AUPRC) score. The AUPRC score gives us the ability of a ranking to place immunogenic neo-peptides before non-immunogenic ones.

(G) Neo-peptide feature importance calculated using Shapley values for LR and XGBoost classifiers trained on NCI-train.

(H) Shapley values of the KTYQGSYGFR neo-peptide (blue bars) from NCI-test patient 4,350 compared with average Shapley values of the top 20 ranked neo-peptides of patient 4,350 (orange bars). LR classifier trained on NCI-train ranked the neo-peptide in rank 3.5 on average. The error bars indicate the standard deviation over the ten replicate runs.

(I) Same as in (H) but for immunogenic neo-peptide DRNIFRHSVV of patient 4,324 in NCI-test (blue bars), which had an average rank of 1,920.1 in the 10 replicate runs.

(J) HLAp length distribution for HLA-C06:02 allele taken from MHC Motif Atlas (<http://mhc motif atlas.org/>).

(K) Bare motif without pseudo-count correction obtained from the 75 HLA-C06:02 10 mers included in HLA Motif Atlas.

See also Figure S4.

classifiers could easily capture inherent biases related to the selection of neo-peptides for screening assays, potentially resulting in suboptimal rankings when applied to other datasets. This justifies our approach of training our classifiers on the NCI dataset, which is characterized by minimal bias, to mitigate the impact of dataset-specific biases and achieve more accurate and reliable rankings.

Next, we compared the performance of our ML ranking methods with an alternative simple approach where neo-peptides were initially sorted based on MixMHCpred or NetMHCpan %rank scores and then by RNA-seq expression to resolve the ties. We demonstrated the superior performance of the ML classifiers compared with this basic ranking strategy (Figures 3A–3C). NetMHCpan performed better than MixMHCpred on the NCI-test and TESLA datasets, where NetMHCpan was used to select neo-peptides for screening, but led to similar ranking for HiTIDE, where MixMHCpred was used for the screening selection. Finally, we compared our results with the rankings published by Gartner et al.<sup>28</sup> for the 23 patients in NCI-test. Our results demonstrated that LR, XGBoost, and the voting classifiers ranked more immunogenic neo-peptides in the top 20, 50, and 100 ranks (Figure 3A; Data S4). Compared with Gartner et al., LR placed 30.0% more neo-peptides into the top 20, 26.7% more into the top 50, and 11.8% more into the top 100.

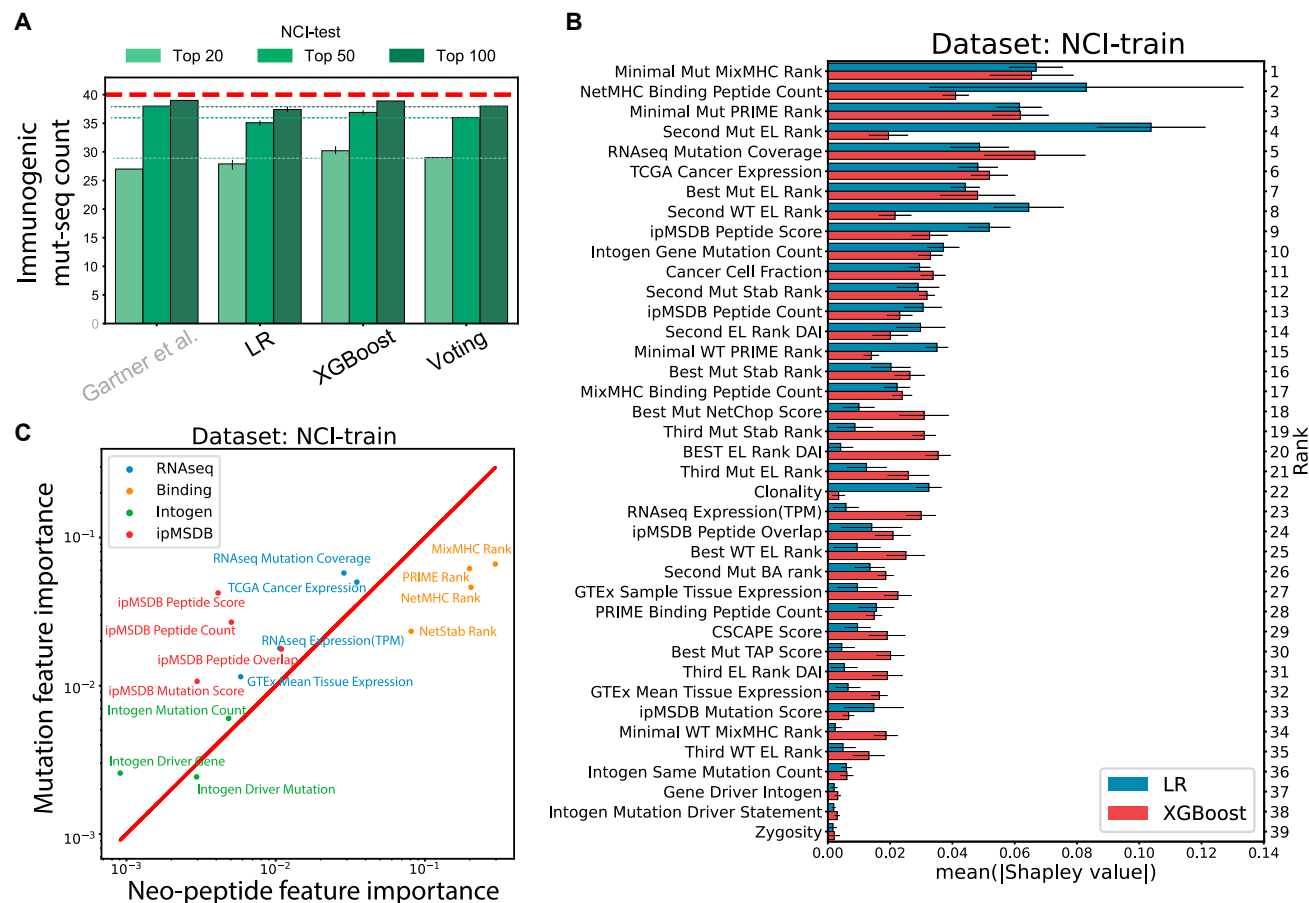
In addition, we conducted a comparison between our ML approach for the TESLA dataset and the consortium results reported by Wells et al. for these data.<sup>26</sup> Our ML ranking achieved the best ranking among the TESLA participants when considering the three evaluation metrics introduced by Wells et al.,<sup>26</sup> with an average rank of 2 compared with the second-best average rank of 3.3 for the “owl” group. Specifically, our voting classifier obtained a median “fraction ranked (FR)” score (see STAR Methods) of 77.8% (Figure 3D), which was better than the FR scores reported by all other groups participating in the TESLA study. Our median “top-20 immunogenic fraction (TTIF)” score of 0.25, was reached by only one other group (Figure 3E), whereas our median area under the precision recall curve (AUPRC) score (0.273) ranked fourth among all participants (Figure 3F). Because the highest-ranking neo-peptides in the lists submitted by the TESLA participants were actually screened in the immunogenicity screens, we here evaluated the TESLA participants partially on their best-ranked peptides, whereas there was no such bias for our ML methods. The results clearly demonstrate that our ML classifiers, trained on the NCI-train dataset, exhibited strong generalization capabilities, and yielded highly accurate results when applied to the independent TESLA dataset.

In order to assess the significance of each feature in the LR and XGBoost rankings, we computed the Shapley values associated with each feature.<sup>50,51</sup> This analysis allowed us to quantify the contribution of each feature in determining the final ranking of neo-peptides by these classifiers. Figure 3G shows that the strongest Shapley values for LR and XGBoost stemmed from MixMHCpred, NetMHCpan, and PRIME rank features, followed by stability rank, TCGA expression and RNA-seq mutation coverage, number of binding HLA alleles, MixMHCpred, DAL, and ipMSDB overlap score. For example, Figure 3H demonstrates the Shapley values for neo-peptide EKIALFQSL of patient 4,350 in NCI-test with an average rank of 48.1 in the ten LR replicates, which is much better than rank 1,641 reported by Gartner

et al. The better ranking resulted from the stronger binding affinity reported by MixMHCpred, PRIME, and NetMHCpan for the HLA-B39:01 allele compared with MHCFlurry v1.6, which was used by Gartner et al., but also IntOGen scores, binding stability, TCGA expression and “ipMSDB Peptide Match Overlap” contributed to the good rank. In contrast, the neo-peptide DRNIFRHSV from patient 4,324 in NCI-test was ranked poorly by our LR classifier (average rank 1,920.1) and by Gartner et al. (rank 24,392) because the peptide had poor %rank scores for allele HLA-C06:02 by all used binding-affinity predictors (Figure 3I), and also Gartner et al. reported a poor %rank with MHCFlurry. The HLA-C06:02 allele binds mainly 9 mers and only a few 10 mers (Figure 3J), resulting in a poor MixMHCpred %rank for 10 mers, even if the 75 10 mer ligands included in the major histocompatibility complex (MHC) Motif Atlas<sup>52</sup> show a clear preference for arginine in the second, and leucine and valine in the 10th position (Figure 3K). Overall ipMSDB and IntOGen features had lower Shapley feature importance, but their contribution to the ranking of immunogenic neo-peptides was still evident (t test p value for rank\_score increase is  $2.54 \times 10^{-8}$  for ipMSDB features, and  $3.70 \times 10^{-10}$  for IntOGen features) (Figure S4L). Excluding these features from LR prioritization reduced the number of neo-peptides ranked in the top 20 in NCI-test by 16.1%.

### Effective ranking of immunogenic mutations requires dedicated training of classifiers

Most neoantigen-based cancer vaccination strategies use long mutated peptides (15–25 mers) or RNA mini-gene constructs encoding such sequences and rely on the selection of the most promising mutations. When prioritizing mutations, the relative importance of mutation features such as RNA-seq expression or coverage is expected to change compared with their significance in prioritization of the minimal neo-peptide sequences (see below). Therefore, instead of using the above neo-peptide classifiers to build a mutation ranking method, we trained mutation classifiers from scratch using the mutation features (Data S3). When we trained the LR and XGBoost classifiers on NCI-train, XGBoost slightly outperformed LR (Figures 4A, S4M, and S4N). For the NCI-test data, both LR and XGBoost performed better in the top 20 than the ranking published by Gartner et al. (Figure 4A; Data S4), but the difference was less pronounced than for neo-peptides. Although binding-affinity features were still most powerful (Figure 4B), the importance of non-HLA-binding-related features, such as RNA-seq coverage, TCGA expression, ipMSDB scores, and IntOGen scores features increased compared with the corresponding neo-peptide features, whereas the importance of binding-affinity ranks decreased (Figure 4C). This emphasizes that prioritizing mutations is different from prioritizing neo-peptides and requires different ML strategies. As for neo-peptides, ipMSDB and IntOGen features contributed complementary information and improved LR based mutation ranking (t test p value for rank\_score increase is 0.0264 for ipMSDB features and 0.0844 for IntOGen features) (Figure S4O). Our approach enabled us to develop specialized classifiers dedicated to mutation and neo-peptide prioritization, thereby ensuring a more tailored and accurate assessment of their importance in the context of neoantigen immunogenicity prediction.



**Figure 4. Effective ranking of immunogenic mutations requires dedicated training of classifiers**

(A) Immunogenic mutations were ranked per patient and the number of immunogenic mutations in the top 20, 50, or 100 ranks was calculated for each patient. The y axis represents these numbers summed over all patients in the dataset. The red dashed horizontal lines indicate the total number of immunogenic mutations in a dataset. The number of top-ranking immunogenic mutations is shown for patients in NCI-test for the LR, XGBoost, and voting classifiers. Gartner et al. refers to the ranking reported in Gartner et al.<sup>26</sup> The horizontal green lines mark the mean performance of the voting classifier in the top 20, 50, or 100 ranks according to their respective colors.

(B) Mutation Shapley feature importance for the LR and XGBoost classifiers in NCI-train. The error bars indicate the standard deviation over 10 replicate runs. The features on the y axis are ordered by decreasing feature importance of both LR and XGBoost.

(C) Neo-peptide feature importance (Figure 3G) compared with mutation feature importance (B) for RNA-seq expression-, binding affinity-, IntOGen-, and ipMSDB-related features used by both neo-peptides and mutation classifiers.

See also Figure S4.

## DISCUSSION

Accurate prediction and prioritization methods of patient-specific neoantigens is still an important barrier for development of effective cancer vaccines and neoantigen-based T cell therapies. Because currently the number of mutations included in a personalized cancer vaccine is in the range of about 20 mutations, the selection of mutations is rather straightforward in case of low tumor mutational burden (TMB)<sup>53</sup>; however, this becomes a critical challenge in the medium to high TMB. Furthermore, the utilization of different validation assays for assessing immunogenicity in various laboratories, along with the use of diverse protocols for T cell isolation and expansion,<sup>54</sup> has a potential to introduce variations, underscoring the importance of harmonizing datasets and providing prediction solutions with generalized good performance across labs. Our systematic

analysis of immunogenic and non-immunogenic neoantigens, demonstrated that many feature scores reflecting processes of the antigen presentation machinery, such as binding affinity and stability, RNA expression and coverage, the presence of non-mutated counterparts of neo-peptides in immunopeptidome hotspots, binding promiscuity, and the role of the mutated gene in oncogenicity, were all predictive for immunogenicity across datasets and immunogenicity validation methods. Indeed, a neoantigen quality model incorporated similar features, such as the differential presentation and T cell cross reactivity against the neoantigen and its WT counterpart.<sup>55</sup> Variations of this model were applied to predict the survival of patients treated with anti-CTLA4 and anti-PD-1,<sup>55</sup> to predict immune editing in long term survivors of pancreatic ductal adenocarcinoma (PDAC),<sup>56</sup> and the induction of neoantigen-specific T cell responses following treatment with personalized mRNA vaccine.<sup>53</sup>



However, the applicability of the “high-quality” model is limited to providing predictions solely for 9-mer peptides and the model does not consider the important information from RNA-seq data. The complex multidimensional structure of the feature manifold motivated the use of ML techniques, to efficiently combine these features for the prioritization of neo-peptides or mutations.

Beyond the selection of the descriptive features, we evaluated several data normalization methods and found that they had a strong impact on the outcome. In addition, we applied the Hyperopt<sup>45</sup> framework to find the optimal classifier hyperparameters, a technical step that is important for the overall performance of ML tools. Several classifier algorithms were then trained on the large NCI<sup>27,28</sup> cohort, which was the least biased and most comprehensive of the three datasets. We observed that the LR<sup>46</sup> and XGBoost<sup>47</sup> classifiers outperformed the others and that their results were to some extent complementary, motivating the use of a voting classifier, which combined the LR and XGBoost probabilities and uniformly provided more robust results. Importantly, the LR and XGBoost classifiers trained on NCI-train resulted in accurate immunogenicity rankings for neoantigens in the TESLA<sup>26</sup> and in-house HiTIDE datasets, which have different HLA restrictions, originate in different tumor types, which were obtained from different laboratories and screened with different immunogenicity assays. Our ML ranking achieved the highest position among the TESLA participants when considering all three evaluation metrics. Additionally, our classifiers surpassed the performance of the classifier reported by Gartner et al. for the NCI-test dataset<sup>28</sup> in which our approach resulted in a remarkable 30% increase in the number of immunogenic neo-peptides ranked within the top 20.

In order to assess the significance of features in the classification task, we used Shapley values.<sup>50,51</sup> For prioritization of mutations, features describing both the mutations (e.g., RNA expression and ipMSDB) and their neo-peptides (e.g., binding affinity) had high importance. In contrast, for prioritization of neo-peptides, binding affinity and stability features dominated. Overall, the performance of HLA binding prediction tools has greatly improved over the last years, especially due to the availability of high-scale accurate MS data of eluted HLA peptides and the implementation of advanced ML approaches. However, our analysis showed that for some peptides that failed to be placed in the top ranks, the limiting factor was, to our surprise, the still suboptimal accuracy of the HLA binding affinity prediction. Nevertheless, we demonstrated that many other features are positively associated with the ranking. This was particularly visible when we excluded ipMSDB and IntOGen features from the features set used for the classification, leading to a decrease in performance.

Our classifiers perform well for datasets with different immunogenicity validation methods, providing an advantage that allows them to be utilized by diverse groups, irrespective of their chosen validation methods. Our results will contribute to immunogenicity prediction in two scenarios. First, users can reproduce all the features we included in our work and apply our trained classifiers directly for antigen prioritization on their data or combine our classifiers with classifiers trained on their own data. Second, our harmonized datasets can serve as a basis. The available features can be edited, and additional features

can be included. Users can train and benchmark their own classifiers and ML methods with these datasets. To conclude, together with our ML classifiers and ML methods, we provide easily accessible data for method development and benchmarking with the aim to improve the selection of immunogenic neo-peptides and mutations for the development of effective personalized immunotherapy treatments.

### Limitations of the study

Of note, some potential limitations should be considered. The datasets may contain some false-negative neo-peptides because only a subset was screened for immunogenicity. It is equally important to note that the assessment of neoantigen-specific responses may underestimate their true potential due to the possibility of T cell exhaustion, which can result in limited expansion or diminished reactivity during *in vitro* culture.<sup>57</sup> This situation could be improved by screening more neo-peptides per mutation or by applying semi-supervised learning methods, which use a combination of clustering and classification algorithms to correct the labels of some wrongly assigned data points. In this study we exclusively considered SNV SMs, but it is known that peptides mapped to insertions, deletions, and out-of-frame and gene fusion events have a high immunogenicity potential due to their increased dissimilarity to WT HLA-bound peptides. However, the amount of immunogenicity data available for non-SNV genomic alterations is limited. To circumvent this limitation, one could leverage the predictors built on SNV mutations and neo-peptides to predict the immunogenicity for non-SNV mutations too. In addition, HLA loss of heterozygosity and expression silencing frequently occurs in cancers<sup>41,58</sup> and such silenced HLA alleles may be excluded from HLA binding and stability predictions. Furthermore, once enough data for CD4<sup>+</sup> T cell recognition of neo-peptides will be available, predictors for neoantigens bound to HLA-II complexes may apply a similar approach.<sup>7,59–63</sup>

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Datasets

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.immuni.2023.09.002>.

### ACKNOWLEDGMENTS

This study was supported by the Ludwig Institute for Cancer Research, by grant KFS-4680-02-2019 from the Swiss Cancer Research Foundation (M.B.-S.), and by the Swiss National Science Foundation, PRIMA grant PR00P3\_193079 (M.B.-S.). This work was also supported by grants from

Cancera, Mats Paulssons, and by a gift from the Biltema Foundation that was administered by the ISREC Foundation, Lausanne, Switzerland. We thank David Gfeller for very insightful discussions and critical remarks.

#### AUTHOR CONTRIBUTIONS

M.M. and M.B.-S. designed the study and drafted the manuscript. F.H., A.I.K., and B.J.S. designed, implemented, and executed the analysis of the WES and RNA-seq data. F.H., M.M., and E.R.A. wrote the software to calculate feature scores. M.M. designed and implemented ML methods, ML data analysis and visualization. J.M. and M.T.-C. performed sample preparation for WES and RNA-seq. M.A., J.C., and A.H. designed and analyzed the immunogenicity assays. B.M., T.G., and A.A. performed the immunogenicity assays. G.C. oversees clinical phase I trials, provided access to patient material, and helped with data interpretation. All authors read and helped revise the paper.

#### DECLARATION OF INTERESTS

The research presented in this paper is associated with the pending patent application PCT/EP2022/082845. The inventors listed on the patent application are M.B.-S., F.H., and M.M.

#### INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: March 27, 2023

Revised: June 26, 2023

Accepted: September 5, 2023

Published: October 9, 2023

#### REFERENCES

- Tran, E., Ahmadzadeh, M., Lu, Y.C., Gros, A., Turcotte, S., Robbins, P.F., Gartner, J.J., Zheng, Z., Li, Y.F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350, 1387–1390. <https://doi.org/10.1126/science.aad1253>.
- Tran, E., Robbins, P.F., Lu, Y.C., Prickett, T.D., Gartner, J.J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E.M., et al. (2016). T-cell transfer therapy targeting mutant KRAS in cancer. *N. Engl. J. Med.* 375, 2255–2262. <https://doi.org/10.1056/NEJMoa1609279>.
- Chen, F., Zou, Z., Du, J., Su, S., Shao, J., Meng, F., Yang, J., Xu, Q., Ding, N., Yang, Y., et al. (2019). Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors. *J. Clin. Invest.* 129, 2056–2070. <https://doi.org/10.1172/JCI99538>.
- Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., et al. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124–128. <https://doi.org/10.1126/science.aaa1348>.
- Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371, 2189–2199. <https://doi.org/10.1056/NEJMoa1406498>.
- Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Löwer, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrörs, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222–226. <https://doi.org/10.1038/nature23003>.
- Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. <https://doi.org/10.1038/nature22991>.
- Schumacher, T.N., and Schreiber, R.D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74. <https://doi.org/10.1126/science.aaa4971>.
- Yarchoan, M., Johnson, B.A., Lutz, E.R., Laheru, D.A., and Jaffee, E.M. (2017). Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* 17, 209–222. <https://doi.org/10.1038/nrc.2016.154>.
- Hadrup, S.R., Bakker, A.H., Shu, C.J., Andersen, R.S., van Veluw, J., Hombrink, P., Castermans, E., Thor Straten, P., Blank, C., Haanen, J.B., et al. (2009). Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat. Methods* 6, 520–526. <https://doi.org/10.1038/nmeth.1345>.
- Bentzen, A.K., Marquard, A.M., Lyngaa, R., Saini, S.K., Ramskov, S., Donia, M., Such, L., Furness, A.J.S., McGranahan, N., Rosenthal, R., et al. (2016). Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* 34, 1037–1045. <https://doi.org/10.1038/nbt.3662>.
- Bentzen, A.K., and Hadrup, S.R. (2017). Evolution of MHC-based technologies used for detection of antigen-responsive T cells. *Cancer Immunol. Immunother.* 66, 657–666. <https://doi.org/10.1007/s00262-017-1971-5>.
- Arnaud, M., Chiffelle, J., Genolet, R., Navarro Rodrigo, B., Perez, M.A.S., Huber, F., Magnin, M., Nguyen-Ngoc, T., Guillaume, P., Baumgaertner, P., et al. (2022). Sensitive identification of neoantigens and cognate TCRs in human solid tumors. *Nat. Biotechnol.* 40, 656–660. <https://doi.org/10.1038/s41587-021-01072-6>.
- Buckley, P.R., Lee, C.H., Ma, R., Woodhouse, I., Woo, J., Tsvetkov, V.O., Shcherbinin, D.S., Antanaviciute, A., Shughay, M., Rei, M., et al. (2022). Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens. *Brief. Bioinform.* 23, bbac141. <https://doi.org/10.1093/bib/bbac141>.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 48, W449–W454. <https://doi.org/10.1093/nar/gkaa379>.
- Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P.O., Kandalaf, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput. Biol.* 13, e1005725. <https://doi.org/10.1371/journal.pcbi.1005725>.
- O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7, 129–132.e4. <https://doi.org/10.1016/j.cels.2018.05.014>.
- Pylke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., McClory, R., et al. (2021). Precision neoantigen discovery using large-scale immunopeptidomes and composite modeling of MHC peptide presentation. *Mol. Cell. Proteomics* 20, 100111. <https://doi.org/10.1016/j.mcpro.2021.100111>.
- Duan, F., Duitama, J., Al Seesi, S., Ayres, C.M., Corcelli, S.A., Pawashe, A.P., Blanchard, T., McMahon, D., Sidney, J., Sette, A., et al. (2014). Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* 211, 2231–2248. <https://doi.org/10.1084/jem.20141308>.
- Koşaloğlu-Yalçın, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premial, A.L.R., Sidney, J., Vaughan, K., Greenbaum, J., Robbins, P., Gartner, J., Sette, A., et al. (2018). Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncol Immunology* 7, e1492508. <https://doi.org/10.1080/2162402X.2018.1492508>.
- Builik-Sullivan, B., Busby, J., Palmer, C.D., Davis, M.J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2018). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37, 55–63. <https://doi.org/10.1038/nbt.4313>.
- Richman, L.P., Vonderheide, R.H., and Rech, A.J. (2019). Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* 9, 375–382.e4. <https://doi.org/10.1016/j.cels.2019.08.009>.

23. Schmidt, J., Smith, A.R., Magnin, M., Racle, J., Devlin, J.R., Bobisse, S., Cesbron, J., Bonnet, V., Carmona, S.J., Huber, F., et al. (2021). Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoeediting. *Cell Rep. Med.* **2**, 100194. <https://doi.org/10.1016/j.xcrm.2021.100194>.
24. Bjerregaard, A.M., Nielsen, M., Hadrup, S.R., Szallasi, Z., and Eklund, A.C. (2017). MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* **66**, 1123–1130. <https://doi.org/10.1007/s00262-017-2001-3>.
25. Hundal, J., Kiwala, S., McMichael, J., Miller, C.A., Xia, H., Wollam, A.T., Liu, C.J., Zhao, S., Feng, Y.Y., Graubert, A.P., et al. (2020). pVACtools: A computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol. Res.* **8**, 409–420. <https://doi.org/10.1158/2326-6066.CIR-19-0401>.
26. Wells, D.K., van Buuren, M.M., Dang, K.K., Hubbard-Lucey, V.M., Sheehan, K.C.F., Campbell, K.M., Lamb, A., Ward, J.P., Sidney, J., Blazquez, A.B., et al. (2020). Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* **183**, 818–834.e13. <https://doi.org/10.1016/j.cell.2020.09.015>.
27. Parkhurst, M.R., Robbins, P.F., Tran, E., Prickett, T.D., Gartner, J.J., Jia, L., Ivey, G., Li, Y.F., El-Gamil, M., Lalani, A., et al. (2019). Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. *Cancer Discov.* **9**, 1022–1035. <https://doi.org/10.1158/2159-8290.CD-18-1494>.
28. Gartner, J.J., Parkhurst, M.R., Gros, A., Tran, E., Jafferji, M.S., Copeland, A., Hanada, K.I., Zacharakis, N., Lalani, A., Krishna, S., et al. (2021). A machine learning model for ranking candidate HLA class I neoantigens based on known neopeptides from multiple human tumor types. *Nat. Cancer* **2**, 563–574. <https://doi.org/10.1038/s43018-021-00197-6>.
29. Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57**, 33–41. <https://doi.org/10.1007/s00251-005-0781-7>.
30. Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* **35**, 2295–2303. <https://doi.org/10.1002/eji.200425811>.
31. Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I.D., Sørensen, M., Nielsen, M., and Buus, S. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* **42**, 1405–1416. <https://doi.org/10.1002/eji.201141774>.
32. Ghorani, E., Rosenthal, R., McGranahan, N., Reading, J.L., Lynch, M., Peggs, K.S., Swanton, C., and Quezada, S.A. (2018). Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* **29**, 271–279. <https://doi.org/10.1093/annonc/mdx687>.
33. Capietto, A.H., Jhunjunwala, S., Pollock, S.B., Lupardus, P., Wong, J., Hänsch, L., Cevallos, J., Chestnut, Y., Fernandez, A., Lounsbury, N., et al. (2020). Mutation position is an important determinant for predicting cancer neoantigens. *J. Exp. Med.* **217**. <https://doi.org/10.1084/jem.20190179>.
34. Chowell, D., Krishna, S., Becker, P.D., Cocita, C., Shu, J., Tan, X., Greenberg, P.D., Klavinskis, L.S., Blattman, J.N., and Anderson, K.S. (2015). TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. USA* **112**, E1754–E1762. <https://doi.org/10.1073/pnas.1500973112>.
35. Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2023). Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Syst.* **14**, 72–83.e5. <https://doi.org/10.1016/j.cels.2022.12.002>.
36. Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017). ‘Hotspots’ of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* **8**, 1367. <https://doi.org/10.3389/fimmu.2017.01367>.
37. Hoyos, D., Zappasodi, R., Schulze, I., Sethna, Z., de Andrade, K.C., Bajorin, D.F., Bandlamudi, C., Callahan, M.K., Funt, S.A., Hadrup, S.R., et al. (2022). Fundamental immune–oncogenicity trade-offs define driver mutation fitness. *Nature* **606**, 172–179. <https://doi.org/10.1038/s41586-022-04696-z>.
38. Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572. <https://doi.org/10.1038/s41568-020-0290-x>.
39. Rogers, M.F., Shihab, H.A., Gaunt, T.R., and Campbell, C. (2017). CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.* **7**, 11597. <https://doi.org/10.1038/s41598-017-11746-4>.
40. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* **46**, 315–326. <https://doi.org/10.1016/j.immuni.2017.02.007>.
41. McGranahan, N., and Swanton, C. (2019). Neoantigen quality, not quantity. *Sci. Transl. Med.* **11**. <https://doi.org/10.1126/scitranslmed.aax7918>.
42. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* **14**, 658–673. <https://doi.org/10.1074/mcp.M114.042812>.
43. Pearson, H., Daouda, T., Granados, D.P., Durette, C., Bonneil, E., Courcelles, M., Rodenbrock, A., Laverdure, J.P., Côté, C., Mader, S., et al. (2016). MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* **126**, 4690–4701. <https://doi.org/10.1172/JCI88590>.
44. Yewdell, J.W., Dersh, D., and Fähræus, R. (2019). Peptide channeling: the key to MHC class I immunosurveillance? *Trends Cell Biol.* **29**, 929–939. <https://doi.org/10.1016/j.tcb.2019.09.004>.
45. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D.D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Comput. Sci. Disc.* **8**, 014008. <https://doi.org/10.1088/1749-4699/8/1/014008>.
46. Cox, D.R. (1958). *The regression analysis of binary sequences*. *J. R. Stat. Soc. B* **20**, 215–232.
47. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16 (Association for Computing Machinery)*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
48. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (Curran Associates, Inc.)*.
49. Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT '92 (ACM)*, pp. 144–152. <https://doi.org/10.1145/130385.130401>.
50. Shapley, L.S. (1953). *A value for n-person games*. In *Contributions to the Theory of Games II (Princeton University Press)*.
51. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.)*.
52. Tadros, D.M., Eggenschwiler, S., Racle, J., and Gfeller, D. (2023). The MHC Motif Atlas: a database of MHC binding specificities and ligands. *Nucleic Acids Res.* **51**, D428–D437. <https://doi.org/10.1093/nar/gkac965>.

53. Rojas, L.A., Sethna, Z., Soares, K.C., Olcese, C., Pang, N., Patterson, E., Lihm, J., Ceglia, N., Guasp, P., Chu, A., et al. (2023). Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* 618, 144–150. <https://doi.org/10.1038/s41586-023-06063-y>.
54. Lim, K.P., and Zainal, N.S. (2021). Monitoring T cells responses mounted by therapeutic cancer vaccines. *Front. Mol. Biosci.* 8, 623475.
55. Łuksza, M., Riaz, N., Makarov, V., Balachandran, V.P., Hellmann, M.D., Solovyov, A., Rizvi, N.A., Merghoub, T., Levine, A.J., Chan, T.A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 551, 517–520. <https://doi.org/10.1038/nature24473>.
56. Łuksza, M., Sethna, Z.M., Rojas, L.A., Lihm, J., Bravi, B., Elhanati, Y., Soares, K., Amisaki, M., Dobrin, A., Hoyos, D., et al. (2022). Neoantigen quality predicts immunoeediting in survivors of pancreatic cancer. *Nature* 606, 389–395. <https://doi.org/10.1038/s41586-022-04735-9>.
57. Wherry, E.J., and Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.* 15, 486–499. <https://doi.org/10.1038/nri3862>.
58. Rosenthal, R., Cadieux, E.L., Salgado, R., Bakir, M.A., Moore, D.A., Hiley, C.T., Lund, T., Tanić, M., Reading, J.L., Joshi, K., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature* 567, 479–485. <https://doi.org/10.1038/s41586-019-1032-7>.
59. Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* 51, 766–779.e17. <https://doi.org/10.1016/j.immuni.2019.08.012>.
60. Alspach, E., Lussier, D.M., Miceli, A.P., Kizhvatov, I., DuPage, M., Luoma, A.M., Meng, W., Lichti, C.F., Esaulova, E., Vomund, A.N., et al. (2019). MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* 574, 696–701. <https://doi.org/10.1038/s41586-019-1671-8>.
61. Kreiter, S., Vormehr, M., van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., Boegel, S., Schrörs, B., Vascotto, F., Castle, J.C., et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692–696. <https://doi.org/10.1038/nature14426>.
62. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* 18, 1009–1021. [https://doi.org/10.1016/S1470-2045\(17\)30516-8](https://doi.org/10.1016/S1470-2045(17)30516-8).
63. Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G., and Serody, J.S. (2019). Alternative tumour-specific antigens. *Nat. Rev. Cancer* 19, 465–478. <https://doi.org/10.1038/s41568-019-0162-4>.
64. Kawaguchi, S., and Matsuda, F. (2020). High-definition genomic analysis of HLA genes via comprehensive HLA allele genotyping. *Methods Mol. Biol.* 2131, 31–38. [https://doi.org/10.1007/978-1-0716-0389-5\\_3](https://doi.org/10.1007/978-1-0716-0389-5_3).
65. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
66. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>.
67. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* 26, 64–70. <https://doi.org/10.1093/annonc/mdl479>.
68. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. <https://doi.org/10.1038/nbt.2514>.
69. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. <https://doi.org/10.1101/gr.129684.111>.
70. Barras, D., Ghisoni, E., Chiffelle, J., Orcurto, A., Dagher, J., Fahr, N., Benedetti, F., Crespo, I., Zimmermann, S., Duran, R., et al. (2022). Tumor microenvironment cellular crosstalk predicts response to adoptive TIL therapy in melanoma. Preprint at bioRxiv. <https://doi.org/10.1101/2022.12.23.519261>.

## Resource

# Integrating inflammatory biomarker analysis and artificial-intelligence-enabled image-based profiling to identify drug targets for intestinal fibrosis

Shan Yu,<sup>1,\*</sup> Alexandr A. Kalinin,<sup>3</sup> Maria D. Paraskevopoulou,<sup>2</sup> Marco Maruggi,<sup>1</sup> Jie Cheng,<sup>2</sup> Jie Tang,<sup>1</sup> Ilknur Icke,<sup>2</sup> Yi Luo,<sup>1</sup> Qun Wei,<sup>1</sup> Dan Scheibe,<sup>1</sup> Joel Hunter,<sup>1</sup> Shantanu Singh,<sup>3</sup> Deborah Nguyen,<sup>1</sup> Anne E. Carpenter,<sup>3</sup> and Shane R. Horman<sup>1,4,\*</sup>

<sup>1</sup>Takeda Development Center Americas, Inc., San Diego, CA 92121, USA

<sup>2</sup>Takeda Development Center Americas, Inc., Cambridge, MA 02142, USA

<sup>3</sup>Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

<sup>4</sup>Lead contact

\*Correspondence: shan.yu@takeda.com (S.Y.), shane.horman@takeda.com (S.R.H.)

<https://doi.org/10.1016/j.chembiol.2023.06.014>

## SUMMARY

Intestinal fibrosis, often caused by inflammatory bowel disease, can lead to intestinal stenosis and obstruction, but there are no approved treatments. Drug discovery has been hindered by the lack of screenable cellular phenotypes. To address this, we used a scalable image-based morphology assay called Cell Painting, augmented with machine learning algorithms, to identify small molecules that could reverse the activated fibrotic phenotype of intestinal myofibroblasts. We then conducted a high-throughput small molecule chemogenomics screen of approximately 5,000 compounds with known targets or mechanisms, which have achieved clinical stage or approval by the FDA. By integrating morphological analyses and AI using pathologically relevant cells and disease-relevant stimuli, we identified several compounds and target classes that are potentially able to treat intestinal fibrosis. This phenotypic screening platform offers significant improvements over conventional methods for identifying a wide range of drug targets.

## INTRODUCTION

Intestinal fibrosis is a pathophysiological mechanism of intestinal tissue repair that leads to the deposition of desmoplastic connective tissue after injury. This process can be triggered by noxious agents, including infections, autoimmune reactions, and physical, chemical, and mechanical injuries. Under normal physiological conditions, intestinal immune components can help to clear foreign pathogens and facilitate tissue repair through canonical wound healing processes. However, fibrogenesis may occur when the immune response is uncontrolled and persistent, or when injuries repeat, resulting in chronic damage.<sup>1,2</sup> Intestinal fibrosis is one of the most common complications of patients who suffer from inflammatory bowel disease (IBD), occurring in approximately 5% of ulcerative colitis (UC) patients and more than 30% of Crohn's disease patients. The prevalence of IBD increased from 0.5% in 2010 to 0.75% in 2022 in Western countries and is projected to reach 1% in 2030.<sup>3,4</sup> Fibrostenotic complications, including stricture formation and subsequent intestinal obstruction, significantly increase morbidity and hospitalization, surgical intervention, and health care costs.<sup>1</sup> Despite advances in the development of therapeutics for treating IBD, including small molecular weight immunomodulators (prednisone, 5-aminosalicylic acid, tofacitinib, and ozanimod), DNA/RNA replication inhibitors (azathioprine, methotrexate, and 6-mercaptopurine), and large molecular weight anti-

inflammatory biologics (anti-TNF $\alpha$ , anti-integrins, and anti-IL-12/IL-23), the high incidence of intestinal strictures and requirement for surgical interventions remain.<sup>5</sup> The lack of effective drug therapies for fibrostenotic IBD represents an increasing and significant unmet medical need.

At a molecular basis, intestinal fibrosis in IBD is a dynamic and multifactorial process. It is a consequence of local chronic inflammation and subsequent activation of fibroblasts. Mucosal inflammation occurs when the mucosal integrity is compromised resulting in the influx of micro-organisms from the gut lumen. Myeloid cells, such as macrophages and dendritic cells, recognize these pathogen-associated molecular patterns via Toll-like and NOD-like pattern recognition receptors and propagate the immune signaling by recruiting other immune cells to clear the offending pathogens by releasing cytokines and chemokines, such as TNF $\alpha$ , IL-1 $\beta$ , IL-36, and Oncostatin-M (OSM).<sup>6</sup> Tissue repair and wound healing occurs in the resolution of the inflammation process after initial inflammatory responses. However, in the context of chronic inflammation, cytokines and chemokines drive the differentiation and activation of fibroblasts and their subsequent production of extracellular matrix (ECM) proteins. When the balance between production and enzymatic degradation of ECM proteins is lost, intestinal fibrosis occurs.<sup>5</sup> TGF $\beta$  is a key cytokine that is produced in response to inflammation, and is a well-known driver of fibrogenesis.<sup>5,7</sup> Numerous studies have been carried out to address TGF $\beta$ -induced



fibrosis.<sup>7–10</sup> However, due to the broad physiological functions, TGF $\beta$  inhibition induces undesirable toxicities, which override its therapeutic benefits.<sup>11</sup> In contrast, inflammation-associated fibroblasts (IAFs), enriched for expression of many genes associated with colitis and fibrosis, represent another paradigm in addressing IBD-related fibrosis.<sup>12,13</sup>

Due to the failure rate of translational efficacy for many clinical candidates for IBD,<sup>14</sup> there is an increased interest in the exploratory phase of drug discovery, to utilize disease-relevant phenotypic screening to provide more confidence to identify drug targets or small molecules.<sup>15–17</sup> However, lead molecules derived from phenotypic screening campaigns may be difficult to follow up due to intrinsic complexities of generating useful structure-activity relationships, and lack of structure-based drug design input, coupled to the difficulties in predicting and successfully navigating mechanism-associated toxicities. Chemogenomic screening utilizes a library of selective small molecules with annotated targets. The benefit of phenotypically profiling compounds with known targets and mechanisms is to assist generation of mechanistic hypotheses that can initiate ensuing target validation studies. Although focused chemogenomics libraries restrict the surveyable mechanistic space, hit molecules identified from such screens can suggest that their targets are amenable to functional pharmacological modulation, thus providing evidence of the druggability of the targets.<sup>17</sup>

Due to practicality and affordability, drug discovery campaigns typically employ one or a few readily interpretable biomarkers, such as secretory or intracellular markers or gene-of-interest-driven reporters that reflect known biology. Recently, significant interest has arisen in the drug discovery industry to capture high-dimensional cellular morphological changes to stimuli and drug treatments by using an image-based profiling with automated microscopy.<sup>18</sup> This unbiased, inexpensive, and scalable image-based method, most often using the Cell Painting assay, combines multiple organelle stains in a robust assay yielding single-cell profiles composed of thousands of features.<sup>18</sup> Integrated into machine learning and data mining, Cell Painting offers the potential to accelerate therapeutic discovery by identifying drug-induced cellular phenotypes, elucidating modes of action, and characterizing drug toxicities.<sup>18</sup>

In this study, we describe a chemogenomic library screen in human intestinal fibroblasts using both disease-relevant biomarkers and Cell Painting readouts to interrogate targeted small molecules that can alleviate the fibrotic phenotype. We identified clinically relevant hits from both assay readouts, though the mechanisms-of-action of hits from each assay represent distinct fibrotic biology. We identified inflammatory response regulators with the biomarker assay, and tissue plasticity, remodeling, fibrosis, and angiogenesis signaling modulators with the Cell Painting assay. The hits were further confirmed and validated in colonic fibroblasts treated with other pro-fibrotic stimuli. With this integrated approach using both high throughput biomarker analysis and artificial intelligence-enabled morphological profiling, we were able to discover a wide spectrum of physiologically and clinically relevant small molecules and targets for intestinal fibrosis. Typically, such high-dimensional datasets require extensive data mining and analysis with trained informatics experts to dissect the information. Here, this study serves as a general roadmap to bench scientists without ma-

chine learning skills to identify targets and hits for other complex and challenging phenotypes and polyetiological disease areas.

## RESULTS

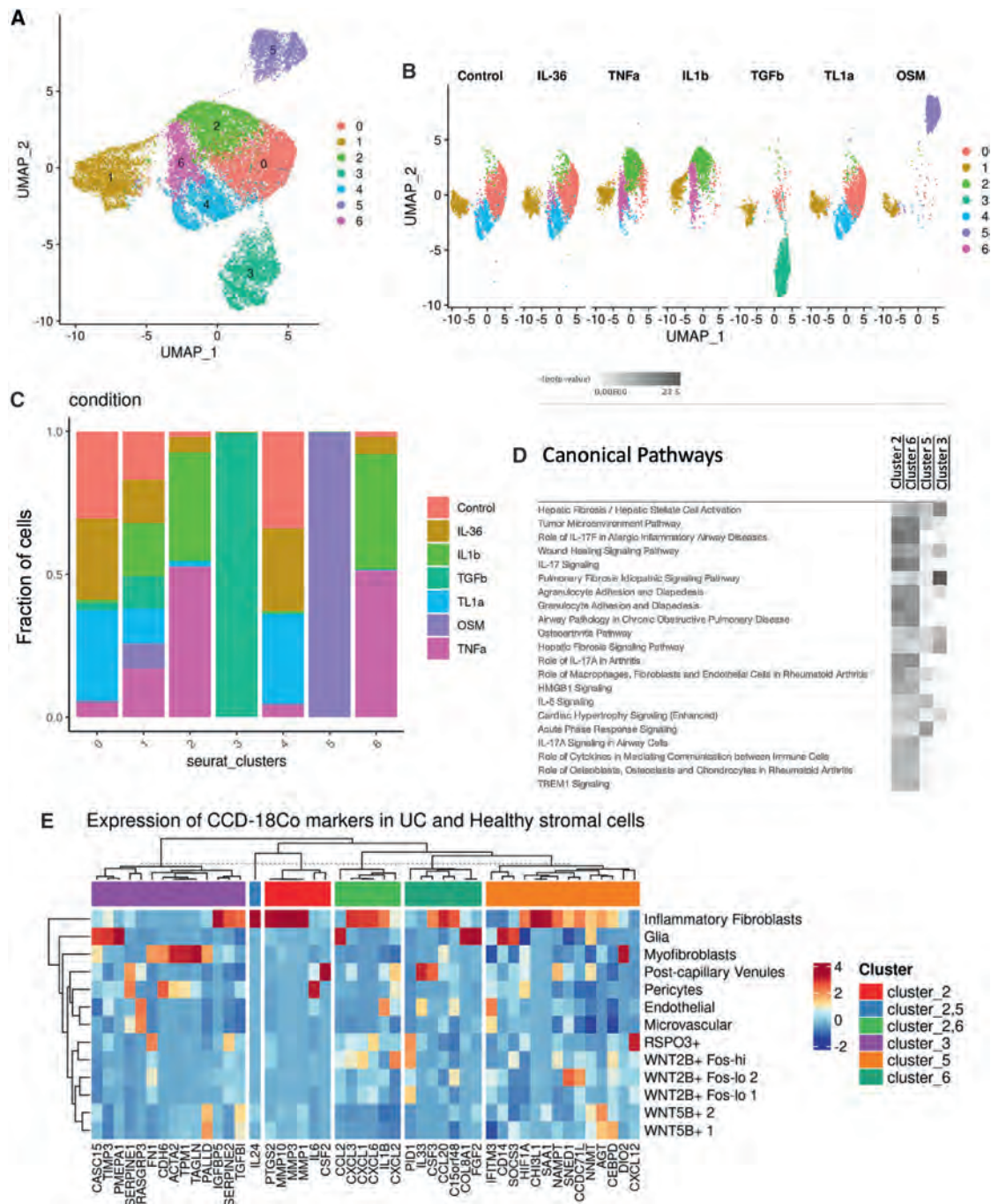
### Development of an *in vitro* cellular disease model that mimics human intestinal fibrosis pathogenic cell population

The CCD-18co human colon fibroblast cell line was identified as a physiologically relevant model for human intestinal fibroblasts.<sup>19</sup> In order to identify culture conditions that yielded the most clinically relevant response to disease-associated stimuli, we performed single-cell transcriptomic analysis of CCD-18co cells that were treated with various pro-fibrotic stimuli, including TNF $\alpha$ , IL-1 $\beta$ , TGF $\beta$ , TL1a, OSM, and IL-36, for 16 h. We combined data from each treatment in an integrated UMAP (Figure 1A) and compared their single-cell RNA sequencing profiles side-by-side (Figure 1B). We identified seven distinct clusters of cells in total, of which several common clusters were shared among all treatments, as well as unique clusters corresponding to particular treatment groups (Figure 1B).

Within these clusters, we performed functional characterization by mapping the enriched canonical pathways and upstream regulators. Clusters 2 and 6 were predominant in TNF $\alpha$  and IL-1 $\beta$  treatment groups (Figure 1C). Genes upregulated in these clusters represented IL-17 signaling, wound healing, TREM1 signaling, cytokine-mediated fibroblast crosstalk, leukocyte migration, and tumor microenvironment pathways; as well as genes involved in mediating inflammatory pathways associated with cancer (Figure 1D). Cluster 3 and cluster 5 were mainly found in TGF $\beta$  and OSM treatments, respectively (Figure 1C). Genes upregulated in cluster 5 represented IL-6 signaling and acute phase response signaling, while genes upregulated in cluster 3 represented tissue fibrosis activities (Figure 1D). IL-36 and TL1A treatment profiles were similar to the control, suggesting neither stimulus exerted a significant effect on the cells (Figure 1B). Upstream regulator detection analysis corroborated that the clusters 2 and 6 are modulated by TNF $\alpha$  and IL-1 $\beta$ , while cluster 3 by TGF $\beta$  and cluster 5 by OSM.

To identify which CCD-18co population exhibited the most disease-mimetic gene expression profile, we mapped activated CCD-18co clusters (clusters 2, 6, 3, and 5) to cell populations from primary human colon stromal biopsies from healthy and UC patients<sup>12</sup> (Figure 1E). We found that clusters 2 and 6, most prevalent in TNF $\alpha$  and IL-1 $\beta$  treatments, and cluster 5, unique to OSM treatment, had signatures that closely overlapped with those of IAFs in diseased human colon biopsies. Cluster 3, specific to TGF $\beta$  treatment, corresponded to both IAFs and myofibroblasts in human colon biopsies. Because IAFs are the immunological hub of multiple signaling pathways that play important roles during the onset of intestinal inflammation and fibrosis,<sup>7</sup> and IAFs are associated with anti-TNF $\alpha$  drug resistance in IBD patients,<sup>12</sup> we sought to address this key unmet medical need for intestinal fibrosis and perform the primary screen with TNF $\alpha$  as stimulus, as it was found to induce an IAF phenotype.

To quantify the effects of TNF $\alpha$  signaling on morphological fibrosis in CCD-18co cells, we knocked out the *TNFRSF1A* and *TNFRSF1B* genes, which encode TNFR1 and TNFR2 (TNF $\alpha$  cell surface receptors), respectively, individually or together using



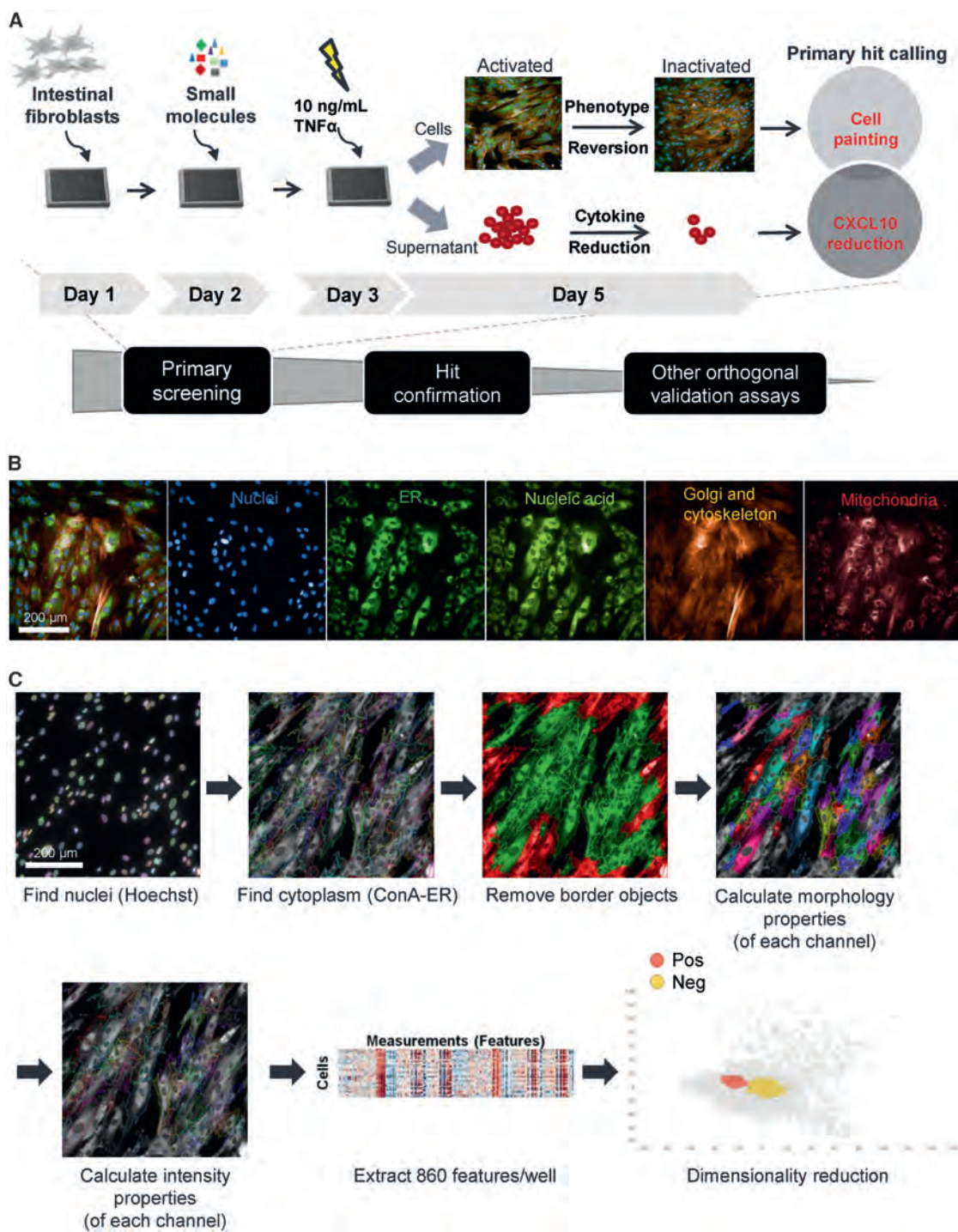
**Figure 1. Bioinformatic analysis of transcriptome profile of CCD-18co cells and comparison with human colon biopsies**

(A and B) UMAP embedding of 16,750 single-cell RNA sequencing (scRNA-seq) profiles from CCD-18co fibroblast cell cultures with different stimuli, including TNF $\alpha$ , IL-1 $\beta$ , TGF $\beta$ , TL1a, OSM, and IL-36, for 16 h. Seven identified single-cell clusters are indicated by colors.

(C) Stacked bar graph showed cluster cell composition, with cluster 3 corresponding to cells activated by TGF $\beta$ , cluster 5 corresponding to cells stimulated by OSM, while clusters 2 and 6 captured cellular responses upon TNF $\alpha$  and IL-1 $\beta$  treatments. The remaining clusters were not overrepresented in any of the conditions and were considered baseline state.

(D) Ingenuity pathway analysis (IPA) canonical pathways associated with the upregulated genes in clusters 2 and 6 (TNF $\alpha$  and IL-1 $\beta$  stimuli), cluster 5 (OSM), and cluster 3 (TGF $\beta$ ). Cluster 1, not shown in Figure 1D, exhibited high expression of cell cycle phase genes. Grayscale represents p-score =  $-\log_{10}$  (p value).

(E) Top 15 markers from CCD-18co fibroblast cell clusters 2, 3, 5, and 6 were analyzed in human colon fibroblasts from UC and healthy patients, retrieved from the published stromal single cell atlas.<sup>12</sup> Highly expressed genes in CCD-18co clusters 2, 5, and 6 (TNF $\alpha$ , OSM, and IL-1 $\beta$  treatments) were enriched in inflammatory fibroblasts, and highly expressed genes in CCD-18co cluster 3 (TGF $\beta$  cellular treatment) were elevated in myofibroblasts from colonic biopsies.



**Figure 2. Chemogenomic library screen workflow**

(A) The screen was conducted through a process including primary screening, hit confirmation, and orthogonal validation assays. For the primary screen, colonic fibroblasts CCD-18co cells were plated on day 1, followed by small molecule transfer on day 2, and 10 ng/mL TNF $\alpha$  stimulation on day 3. The supernatant samples were collected for the CXCL10 reduction assay and cells were stained with the Cell Painting dyes for the high content imaging assay. Hits from both assays were called and analyzed individually and collectively.

(B) CCD-18co cells that were stained with Cell Painting dyes including Hoechst 33342 (nuclei), Concanavalin A-Alexa 488 (ER), SYTO 14 (nucleic acid), WGA-Alexa 555 (Golgi), phalloidin-Alexa 568 (cytoskeleton) and MitoTracker Deep Red (mitochondria), and imaged with Operetta CLS. The image on the far left represents the merged image of all channels.

(legend continued on next page)



CRISPR/Cas9 gene editing (Figure S1A, related to Figures 1 and 2), then evaluated the response of the cells to TNF $\alpha$ . Upon activation of NF- $\kappa$ B by TNF $\alpha$  signaling, p65, a subunit of NF- $\kappa$ B also known as RELA, was observed to translocate from the cytoplasm to the nucleus (Figure S1B, related to Figures 1 and 2). However, cells transfected with individual or pooled TNFRSF1A guide RNAs (gRNAs) showed that p65 remained, at least partially, in the cytoplasm (Figure S1C, related to Figures 1 and 2), indicating reduced NF- $\kappa$ B signaling. Further, CCD-18co cells transfected with individual or pooled TNFRSF1A gRNAs showed a trend toward diminished CXCL10 secretion compared to control cells (Figure standard deviations (S.D.) related to Figures 1 and 2). The effect of dual TNFRSF1A and TNFRSF1B knockout was similar to TNFRSF1A knockout alone indicating TNF $\alpha$  signaling was mediated, at least partially, through TNFR1 instead of TNFR2 in CCD-18co cells.

In high-throughput screening, it is important to use clinically proximal readouts whenever possible to ensure the observed phenotype is a robust surrogate for disease pathology. To that end, we assessed protein and mRNA expression levels of a panel of inflammation-related biomarkers in CCD-18co cells that were treated with disease-relevant pro-fibrotic stimuli. We identified CXCL10 as a significantly upregulated biomarker at both protein and mRNA levels by multiple stimuli, including TNF $\alpha$ , IL-1 $\beta$ , and IL-36 (Figure S2, related to Figure 2). Because CXCL10 contributes to fibrosis by supporting monocyte/macrophage recruitment, angiogenesis, fibroblast collagen synthesis, myofibroblast activation and differentiation, and modulation of CXCL10 and its receptor CXCR3 has been reported to be associated with inflammatory signaling-driven fibrogenesis,<sup>20–24</sup> we chose it as a readout for efficacy in the ensuing screen. Though we did profile more conventional biomarkers of fibrosis, including ACTA2 and COL1A1, neither was induced by pro-fibrotic stimuli at either protein or mRNA level to yield an acceptable assay window for a high throughput screen (Figure S3, related to Figure 2). This is likely due to the fact that they are biomarkers of canonical TGF $\beta$  signaling instead of other pro-inflammatory stimuli (e.g., TNF $\alpha$ , IL-1 $\beta$ , and IL-36).

In addition to CXCL10 secretion as a readout for efficacy, we also used the Cell Painting assay to serve as a morphological readout of cellular fibrosis. Morphologies of CCD-18co cells treated with different pro-fibrotic stimuli were visually distinct (Figure S4A, related to Figure 2) and this translated to cellular features that yielded equally distinct principal component analysis (PCA) plots (Figure S4B, related to Figure 2). Interestingly, the Cell Painting PCA plot strongly resembled the transcriptomic PCA plot (Figure S4C, related to Figures 1 and 2), suggesting CCD-18co cellular morphology might be tightly correlated with gene expression and subsequent biological activities.

### Automated high throughput chemogenomic library screen to identify targeted perturbagens of intestinal fibrosis

To comprehensively profile diverse biological and functional space (Figure 2A), we sourced two small molecule libraries

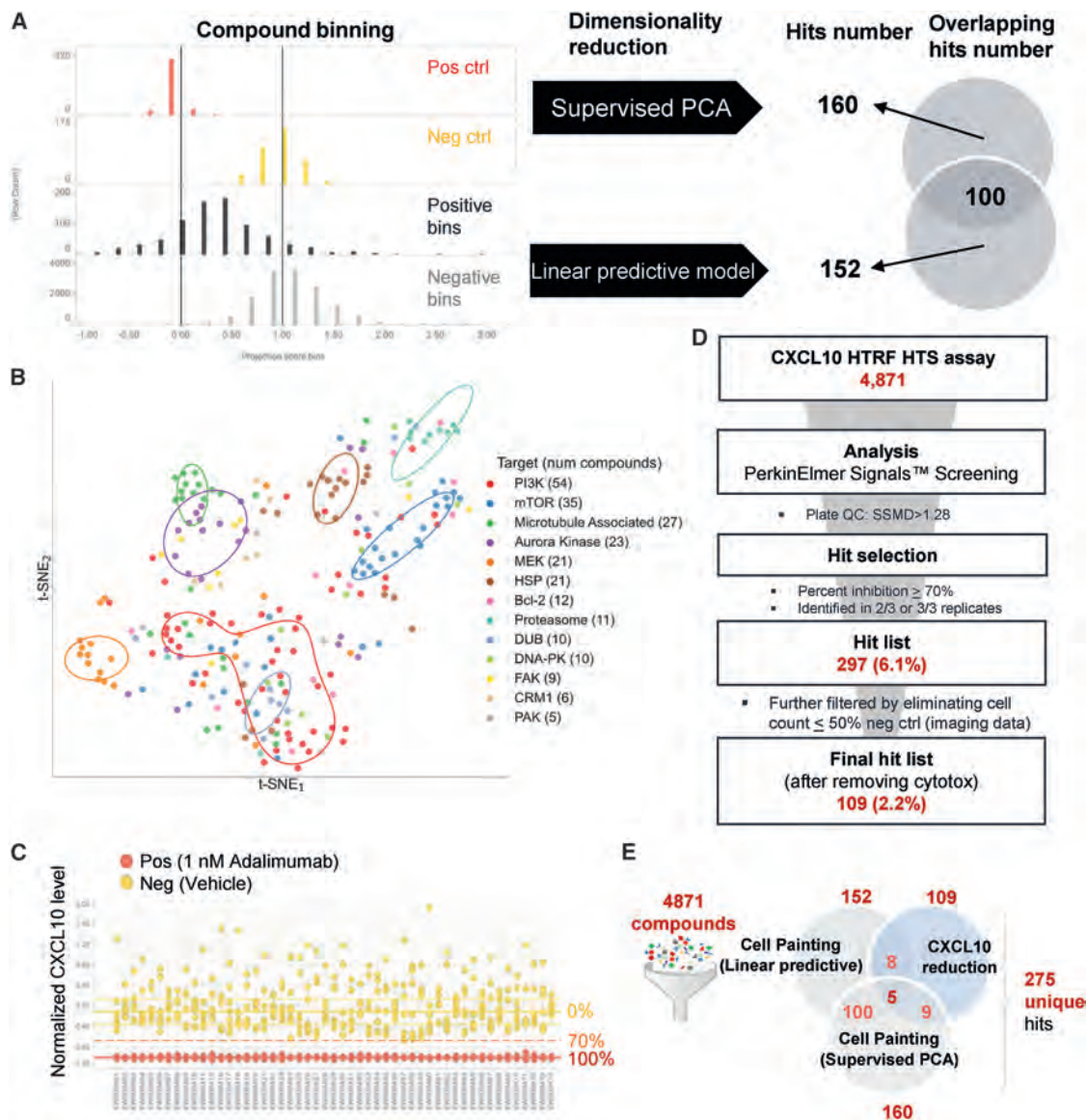
totaling 4,871 compounds annotated with either their reported targets and/or mechanisms of action and have been either tested in clinical trials or approved by the FDA (Selleckchem; Figure S5A, related to Figure 2). The molecular weight and ALogP of these compounds were within the standard range for “drug-like” molecules (Figure S5B, related to Figure 2).<sup>25</sup>

For the primary high throughput screening assay, 1,200 CCD-18co cells/well were plated on the first day, followed by compounds and controls after 24 h (Figure S5C, related to Figure 2). Each compound was tested at 3  $\mu$ M in biological triplicate. 1 ng/mL anti-TNF $\alpha$  antibody adalimumab was used as the positive control, because adalimumab was able to effectively suppress TNF $\alpha$  signaling in the CXCL10 assay (as well as in the Cell Painting assay, as discussed later, Figure S.D. related to Figures 2 and 3). Cells were then treated with 10 ng/mL TNF $\alpha$  on the third day for 48 h, after which time the cell culture supernatants were collected for CXCL10 protein quantitation using a homogeneous time-resolved fluorescence (HTRF) assay.

For the Cell Painting assay, cells from the exact same samples were stained with Cell Painting dyes followed by high-content image acquisition and analysis. The assay includes six fluorescent dyes to highlight different organelles of CCD-18co cells, including MitoTracker Deep Red FM for mitochondria, Concanavalin A-Alexa 488 for endoplasmic reticulum, SYTO 14 for nucleoli and cytoplasmic RNA, WGA-Alexa 555 and phalloidin-Alexa 568 for F-actin cytoskeleton, Golgi, and plasma membrane, Hoechst 33342 for nucleus<sup>26</sup> (Figure 2B). High-content images were captured and cellular morphological features were extracted and then analyzed using a dimensionality reduction method. Compounds that clustered around the positive controls were categorized as Cell Painting hits (Figure 2C). For dimensionality reduction, we used either a supervised PCA or a linear predictive model. For both methods, the medians of positive controls and negative controls were normalized to 0 and 1, respectively. Compounds were then binned into positive or negative bins depending on the projection scores (STAR Methods, Figure 3A, left). Compounds positive for >2 out of 3 replicates in the positive bins and projection scores within the range of Average(pos ctrl)  $\pm$  3  $\times$  S.D. were picked as preliminary hits. Compounds exhibiting cytotoxic profiles were then further filtered based on cell count.

In total, 160 and 152 compounds were picked as hits from supervised PCA and linear predictive models of the Cell Painting data, respectively (Figure 3A, right). There were 100 hits that overlapped between both models for Cell Painting analysis (Figure 3A, right), suggesting the two analytical methods yielded mainly convergent results. In addition, we assessed three other metrics for picking Cell Painting hits; namely using the top 50 features, top 5 features, or top 3 features per channel that separate positive and negative controls, though the hits and targets that were identified were mostly similar (Figure S6, related to Figure 4). To determine whether these cellular features correlate with their biological functions, we projected the cellular features of the targets that were most distinct from the negative controls onto a two-dimensional t-SNE map (Figure 3B). This

(C) Workflow of cellular compartment segmentation of high content images using PerkinElmer Harmony software. Nuclei were identified by Hoechst 33342 stain. Cytoplasm was then identified by Concanavalin A-Alexa 488 stain. The border objects were excluded from analysis. Different morphology and intensity properties of each channel were calculated and 860 features were extracted at the well-level. The profiling dataset was then analyzed with a dimensionality reduction method, such as PCA.



**Figure 3. Primary screen hit picking strategies for the CXCL10 reduction assay and Cell Painting assay**

(A) The Cell Painting dataset was analyzed with both supervised PCA and linear predictive model methods. Projection scores of Cell Painting controls and samples help to determine the similarities between compounds and controls. Compounds in positive bins in the range between Average(projection score)  $\pm$  3 x S.D. were picked as hits.

(B) t-SNE plot shows the phenotypic space of top compound target categories that are farthest from the negative controls.

(C) Pos and neg ctrl data points of CXCL10 HTRF assay. X axis shows the plate barcode, y axis shows the normalized CXCL10 level. Solid yellow line shows 0% inhibition representing the median of the neg ctrl (vehicle), and solid red line shows 100% inhibition representing the median of pos ctrl (1 nM adalimumab). Dotted orange line shows 70% cutoff for hit picking.

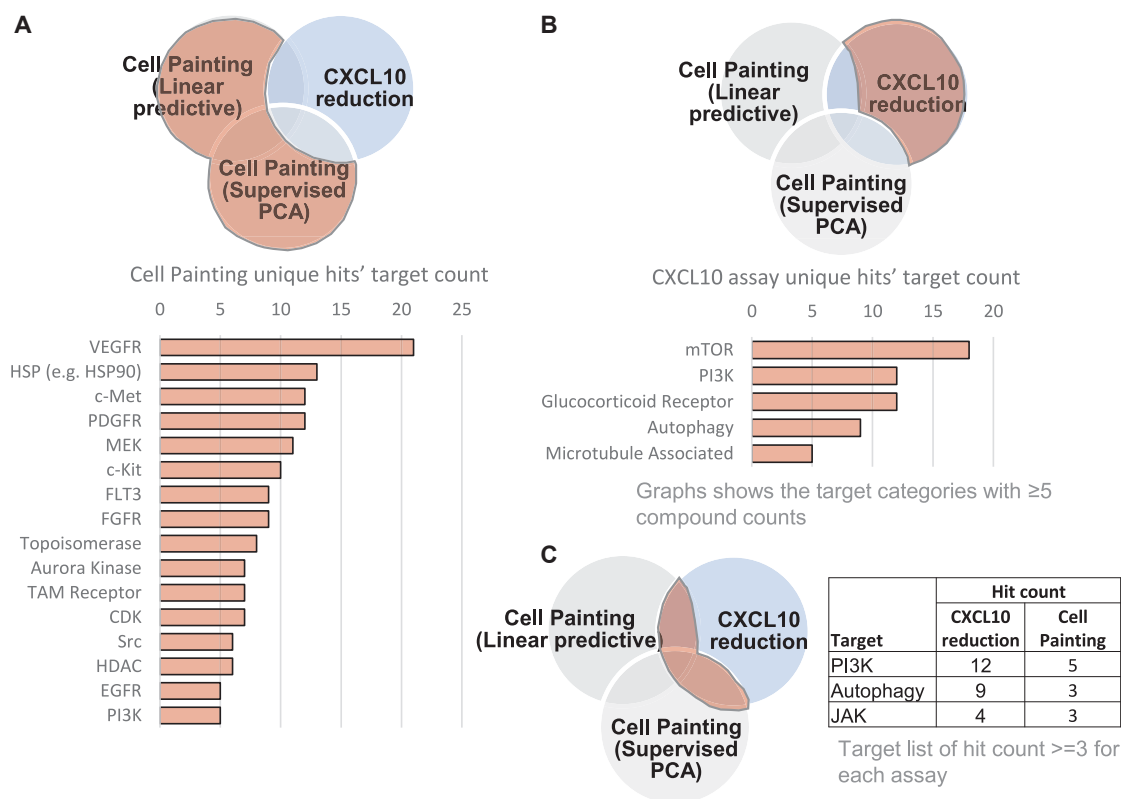
(D) The CXCL10 HTRF assay screening funnel.

(E) Overview of small molecule hit numbers from each assay/analysis.

map showed that some co-annotated compounds form coherent clusters (e.g., MEK and HSP) in phenotypic space whereas others do not (e.g., Bcl-2, FAK, CRM1, and DNA-PK).

For the CXCL10 assay, luminescence intensities of positive and negative controls of each plate were fit on a 0 to 1 scale and were then normalized for their percent inhibition, with the mean of positive control being 100% and the mean of negative control being 0% (Figure 3C). The strictly standardized mean difference

(SSMD) was used to measure the effect size and gauge the assay quality.<sup>27</sup> Plates with SSMD > 1.28 (the SSMD quality cutoff) then proceeded to hit selection. Compounds positive for >2 out of 3 replicates with CXCL10 inhibition >70% were identified as preliminary hits, and then filtered by eliminating cytotoxic compounds (dependent on cell count). After applying this gating strategy, 109 compounds were identified, resulting in a 2.2% hit rate for the CXCL10 screen (Figure 3D).



**Figure 4. Hit category analysis of Cell Painting and CXCL10 reduction assays**

(A) Hit number of target categories for linear predictive model and supervised PCA analysis of Cell Painting. Bar chart shows the target categories with  $\geq 5$  compounds in each.

(B) Hit number of target categories for CXCL10 reduction assay. Bar chart shows the target categories with  $\geq 5$  compounds in each.

(C) Hit number of target categories for the overlapping hits between CXCL10 reduction assay and Cell Painting assays. Table shows target categories with  $\geq 3$  compounds in each assay.

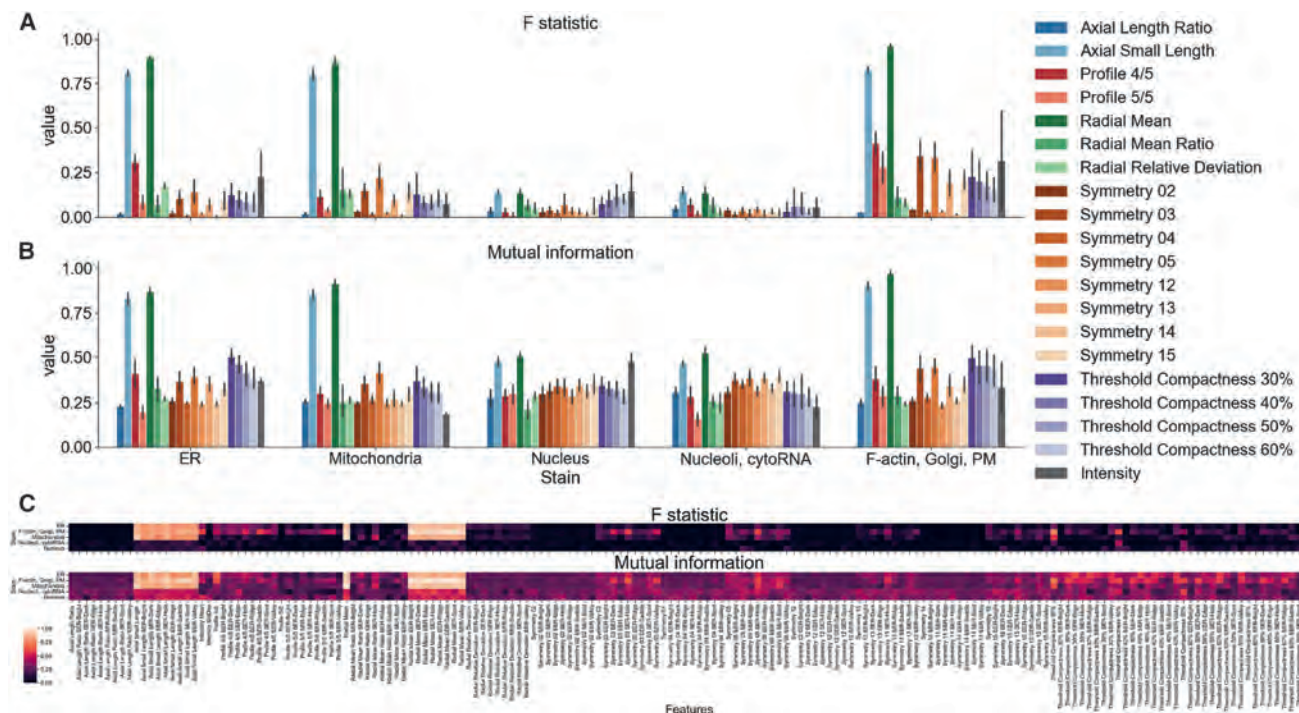
Surprisingly, there were only 8 hits that overlapped between the Cell Painting linear predictive model and the CXCL10 assay, and only 9 hits that overlapped between the Cell Painting supervised PCA model and the CXCL10 assay. Only 5 hits overlapped among all three methods. In the end, after removing duplicate compounds, 275 unique hits from either Cell Painting or CXCL10 assay were advanced for further confirmation and validation (Figure 3E).

### Target discovery through the integration of cytokine biomarker and morphological profiles

It was intriguing that the CXCL10 assay and the Cell Painting assay identified vastly different pools of hit compounds. For hit compounds that were unique to Cell Painting, the top targets included VEGFR, HSP, c-Met and PDGFR, MEK, c-Kit, FLT3, and FGFR (Figure 4A); while hits that were unique to the CXCL10 assay included the targets mTOR, PI3K, glucocorticoid receptor, and several components of the autophagy and microtubule pathways (Figure 4B). For hits that were shared between the two assays, the top targets included PI3K, autophagy, and Janus kinase (JAK) (Figure 4C).

In several contexts, image-based profiles have proven to show predictive abilities for other assays.<sup>28</sup> We wondered whether any particular cellular morphology features from the

Cell Painting assay could be used to predict CCD-18co cells' response to  $\text{TNF}\alpha$ , in terms of secreting CXCL10. We studied the statistical dependence between CXCL10 levels and each of the 860 individual cellular features. Overall, 752 out of 860 features had some linear relationship with the CXCL10 level (F-test,  $p < 0.01$ , Bonferroni-corrected with  $\alpha = 0.01$ ). In particular, we found that a few categories of cellular features including axial small length (the length of the cell's shorter axis in pixel units) and Radial Mean (the mean object radius based on the intensity values weighted by the distance from the mass center) from the ER, mitochondria and F-actin, Golgi and PM channels ( $n = 54$  features) had strong relationships with CXCL10, as indicated by higher average F statistic values (97<sup>th</sup> percentile of distribution of F statistic, all adjusted p values = 0.0), which capture the linear dependency between features and the CXCL10 (Figure 5A). We further confirmed this finding by also calculating averaged mutual information (MI), which is a nonparametric measure that can capture any kind of statistical dependency, and demonstrated that these feature categories have strongest relationships with the CXCL10 level (98<sup>th</sup> percentile of distribution of MI values) (Figure 5B). To focus on the subcategories and examine which particular features had the strongest statistical dependency with CXCL10 level, we found that several Radial Mean features including Edge, Ridge, and Spot of the



**Figure 5. Correlation analysis of morphological features with CXCL10 level**

(A and B) F statistic (which shows the linear dependency) and mutual information (which shows any type of dependency, including linear dependency) between cellular feature subcategories and the CXCL10 level. Error bars show a bootstrap-estimated 95% confidence interval.

(C) Heatmap of top highly correlated features of each subcategory with CXCL10.

Spots, Edges and Ridges (SER) texture analysis in the F-actin, Golgi, and plasma membrane channel have nearly perfect statistical dependency with the CXCL10 level (e.g., Radial Mean SER-Spot has F-statistic of 1.0 and Radial Mean SER-Edge has MI of 1.0) (Figure 5C); indicating these features have strong dependency with CXCL10 and can be considered as potential predictors of CXCL10 level.

### Target validation using pro-fibrotic stimuli-treated cell models

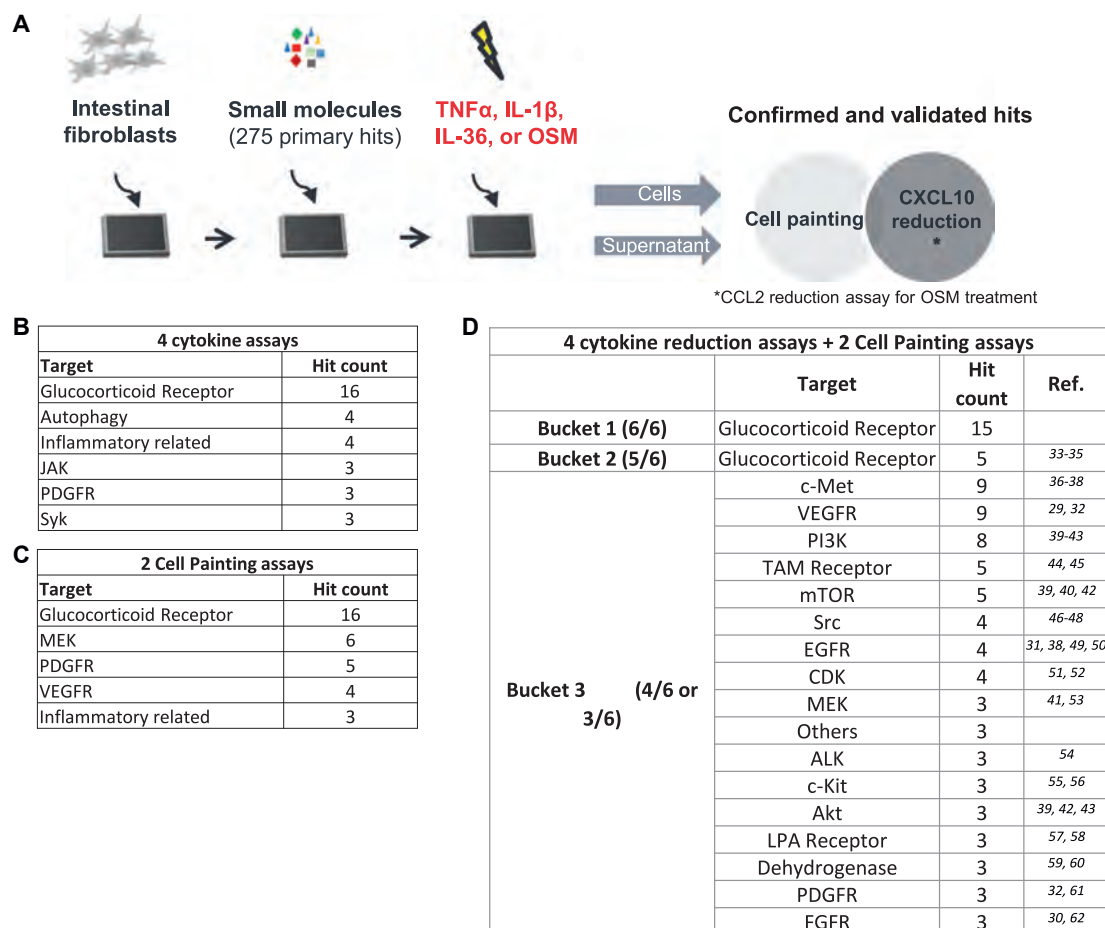
To further characterize hit compounds according to their ability to ameliorate fibrosis from pro-fibrotic stimuli other than TNF $\alpha$ , we profiled the 275 unique hit compounds at three doses (3  $\mu$ M, 0.6  $\mu$ M, and 0.125  $\mu$ M) in assays with different stimuli (IL-1 $\beta$ , IL-36, or OSM) in addition to TNF $\alpha$  (Figure 6A). The CXCL10 assay was used for TNF $\alpha$ -, IL-1 $\beta$ -, and IL-36-treated cells, while a CCL2 assay conducted 2 h post-OSM treatment was used for OSM-treated cells, because CCL2 (Figure S.D. related to Figure 6) but not CXCL10 (Figure S.D. related to Figure 6) is a functional biomarker for OSM stimulation. Similar to CXCL10, CCL2 contributes to fibrosis by recruiting monocyte/macrophage and myofibroblast activation and differentiation.<sup>23</sup> The Cell Painting assay was only used for TNF $\alpha$  and IL-1 $\beta$  stimulation, as there were no viable assay windows for cells treated with either IL-36 or OSM (Figure S4B, related to Figure 6), leaving four cytokine assay and two Cell Painting assay results available for analysis.

The TNF $\alpha$ -stimulated reconfirmation screen of 275 unique hit compounds yielded a 51% reconfirmation rate for reducing

CXCL10 expression/secretion and a 47% reconfirmation rate for Cell Painting, suggesting the robustness of the primary screening assays (confirmed and validated hit results are shown in Table S1, and details of example hits are shown in Figure S7, related to Figure 6). Using a combinatory approach to examine the target categories, we pooled the four cytokine stimulation results and identified glucocorticoid receptor as the top target with 16 hits. This was followed by autophagy, inflammatory-related mechanisms, JAK, PDGFR, and SYK (Figure 6B). The two Cell Painting reconfirmation assays (TNF $\alpha$  and IL-1 $\beta$ ) similarly showed glucocorticoid receptor to be the top target, followed by MEK, PDGFR, VEGFR, and inflammatory-related mechanisms (Figure 6C).

When considering all six compound lists, the hits were binned into three buckets depending on the number of assays in which they were identified as hits. Bucket one included compounds that were picked as hits in six out of six assays. All hits in this bucket were glucocorticoid receptor modulators (steroids). Bucket two included compounds that were picked as hits in five out of six assays and similarly, all hits in bucket two were mainly glucocorticoid receptor modulators. Bucket three included compounds that were identified as hits in three or four out of six assays and this bucket represented the largest variety of biological functions with different mechanisms of action (Figure 6D).

To understand these targets in the context of signaling pathways, we mined the literature and identified any associations between targets in bucket three and intestinal fibrosis. Overall, three main pathways were identified: ER stress response,



**Figure 6. Hit confirmation and validation assay workflow and hit categories**

(A) Hit confirmation and validation experimental workflow.

(B) Top target categories across the four cytokine reduction assays. Table shows target categories with  $\geq 3$  compounds for each assay.

(C) Top target categories for the Cell Painting results of  $TNF\alpha$  and  $IL-1\beta$  stimulation.

(D) Top target categories for all six assay results. The results were further bucketed into three categories. Bucket 1 includes compounds that showed effects in all 6 assays. Bucket 2 includes compounds that showed effects in 5 out of 6 assays. Bucket 3 includes compounds that showed effects in 3 or 4 out of 6 assays. Table only shows target categories with  $\geq 3$  compounds in each.<sup>29-62</sup>

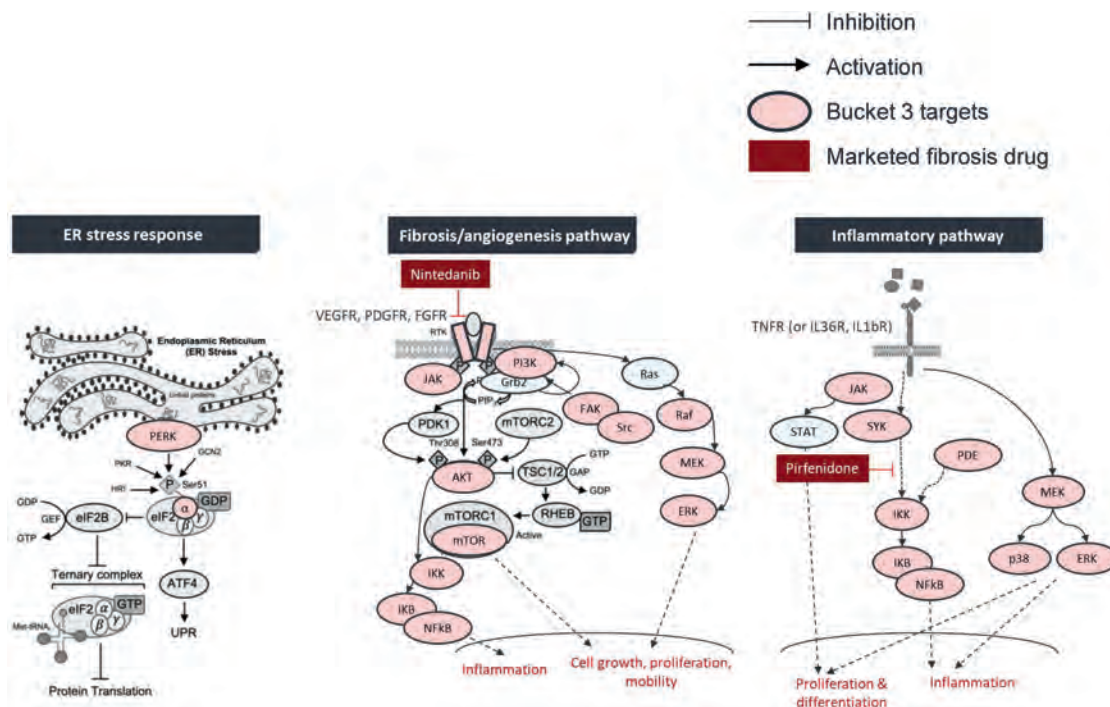
fibrosis/angiogenesis, and inflammation (Figure 7). All three pathways were shown to play a role in tissue fibrosis.<sup>7,9,29-31,63-66</sup> Interestingly, we identified and confirmed both nintedanib (targets PDGFR, VEGFR, and FGFR) and pirfenidone (targets NF- $\kappa$ B), approved drugs for treating idiopathic pulmonary fibrosis (IPF),<sup>32,66</sup> as potent antagonists of myofibroblast activation<sup>67</sup> (Figure 7). These data suggest that the small molecules, targets, and signaling pathways identified through our multi-parametric biomarker and cellular feature profiling approach were physiologically and clinically relevant. Further, this screening platform was able to identify molecules from a wide spectrum of mechanisms of action.

## DISCUSSION

IBD-associated intestinal fibrosis represents a highly invasive and deleterious disease that currently has no approved pharmacological intervention. In order to address this, we developed a clinically relevant humanized intestinal fibrosis model composed

of  $TNF\alpha$ -activated colon fibroblasts. In order to leverage large collections of small molecules for therapeutic profiling efforts, we miniaturized the human IBD fibrosis model to accommodate a scalable phenotypic screening platform for fully automated drug discovery. Employing transcriptomics as a surrogate characteristic for comparing our CCD-18co *in vitro* model to IBD patient biopsies, we identified several distinct transcriptional clusters corresponding to different pro-inflammatory cytokine stimuli. Although  $TGF\beta$  treatment of CCD-18co cells produced a canonical gene expression profile that overlapped with myofibroblast components of patient biopsies, discovery of therapeutics targeting the  $TGF\beta$  pathway has not yielded any clinical treatment due to undesirable toxicities. In recent years, IAFs have been shown to be critical to fibrogenesis associated with chronic inflammatory diseases.<sup>12,13,22</sup> Here, we intended to identify potential therapeutics by targeting IAFs.

As intestinal fibrosis is a result of a complex interplay of immune-mediated inflammatory processes as well as modulation of pro-inflammatory cytokine-mediated signaling pathways,



**Figure 7. Major pathways of the bucket 3 compound targets**

Three major pathways, including ER stress response, fibrosis/angiogenesis pathway, and inflammatory pathway were identified by analyzing the targets of bucket 3 compounds. Pink bubbles show the targets that were identified in the bucket 3 compounds. Gray bubbles show other intermediate targets in the pathway. Nintedanib, a marketed drug for idiopathic pulmonary fibrosis, was identified as a hit in the screen. The screen also identified inflammatory pathway targets through which pirfenidone, another marketed drug for idiopathic pulmonary fibrosis, exerts its effect.

our screening platform required a sophisticated series of assay readouts to account for these polyetiological causes. We first chose to use CXCL10 (IP10) as the primary screen readout due to its well-characterized association with intestinal fibrotic pathology and because compared to other biomarkers, both its mRNA and protein levels were significantly increased by multiple pro-fibrotic stimuli (Figure S2, related to Figure 2). However, to fully assess changes in the fibrotic morphological phenotype, we applied an unbiased image-based profiling technique called Cell Painting. Although Cell Painting has not been widely adapted in the drug discovery industry as a phenotypic readout for efficacy, its scalable ease of use as well as its ability to quantitate changes in thousands of cellular features makes it an ideal method for studying complex biology such as intestinal fibrosis. Cell Painting produces vast morphological information as a collection of extracted cellular features, but by integrating artificial intelligence analytical methods, such as machine learning, we can mine these data to reveal important biological activities of potentially therapeutic small molecules.<sup>18</sup> For example, we found that the relative positions of pro-fibrotic stimuli-treated clusters to vehicle controls in Cell Painting PCA plots were similar to those from RNA-seq PCA plots, suggesting transcriptome profiles and related biological activities strongly correlate with cellular morphological profiles. We also examined whether any specific cellular features were highly correlated with CXCL10 level, because these features may potentially be used as sentinel readouts for CXCL10 in future studies. We identified several subcategories of features, such as Axial Small Length

and Radial Mean in ER, mitochondria and F-actin, Golgi and plasma membrane channels that had high correlations with CXCL10 level (Figure 5).

Surprisingly, we observed divergent hit distribution profiles between the CXCL10 and Cell Painting assay readouts. The reason might be attributed to the fact that only a few cellular features from the Cell Painting assay had a strong statistical correlation to the CXCL10 level (Figure 5). Different cellular features were chosen that better represented the TNF $\alpha$ -stimulated phenotype though they had a lower correlative relationship with CXCL10. These features were chosen for Cell Painting hit selection because they were more prominent in differentiating TNF $\alpha$ -treated and non-treated cells. While the CXCL10 readout identified well-characterized regulators of fibrosis such as mTOR and glucocorticoid receptor, the targets identified through the Cell Painting readout were mechanistically more diverse (e.g., VEGFR, PDGFR, FGFR, c-Met, c-Kit, and MEK) and included such cellular processes as fibrosis, tissue plasticity and remodeling, and angiogenesis. In short, the CXCL10 assay conferred a confidence metric to the biological relevance of our assay platform by identifying several steroid molecules as alleviators of the fibrotic phenotype. However, the Cell Painting assay was able to reveal a diverse array of potential mediators implicated in intestinal fibrosis pathology, expanding the scope of actionable targets. Overall, this high-throughput screening platform combining CXCL10 and Cell Painting readouts was able to identify small molecule hits with proven clinical relevance. For example, our screen identified and confirmed nintedanib, a

drug for treating IPF, may be repurposed to treat intestinal fibrosis. We also identified small molecules that modulate other known fibrosis targets (Figure 6). This suggests the screening platform may be used for repurposing approved or clinical-stage drugs or discovering novel small molecules for intestinal fibrosis.

As our collective understanding of the causes and mediators of disease biology increase, so must our ability to interrogate those causes to discover the next generation of small molecule therapeutics. A complex image-based profiling technique like Cell Painting integrated with state-of-the-art machine learning algorithms to translate thousands of cellular features into disease-relevant targets and pathways may represent a giant leap forward in industrialized drug discovery. Although it may be unlikely that image-based profiling will completely replace conventional biochemical, transcriptional, or proteomic profiling methods, when incorporated into exploratory phases of the drug discovery pipeline, Cell Painting may accelerate the identification of novel therapeutics and expand the targeting space of polyetiological and poorly understood diseases like intestinal fibrosis.

### Limitations of the study

In this study, we utilized CXCL10 and CCL2 as functional readouts for CCD-18co cells due to their robust response to pro-fibrotic stimuli, resulting in an up-regulation of mRNA and protein expression levels (Figure S2B, related to Figure 2). The evidence suggests that CXCL10 and CCL2 play a role in fibrosis by supporting monocyte/macrophage inflammatory response, angiogenesis, fibroblast collagen synthesis, myofibroblast differentiation, and fibroblast recruitment and survival.<sup>23,24</sup> However, it should be noted that the role of CCL2 in fibrosis is somewhat controversial, as there have been reports of CCL2 mediating anti-fibrotic effects in human fibroblasts independently of CCR2.<sup>68</sup> Because of the complexity of intestinal fibrosis and translatability and feasibility of using other validated biomarkers in the cellular screening system, we selected CXCL10 and CCL2 as functional readouts in our screen.

### SIGNIFICANCE

**Our study showed that the integration of Cell Painting morphological profiling with biomarker analysis can be used to identify potential targets and small molecule drugs for a broad spectrum of polyetiological and poorly understood diseases, such as intestinal fibrosis. Here, we provide a roadmap for bench scientists without sophisticated informatics tools or machine learning skills to analyze high dimensional Cell Painting datasets and incorporate image-based profiling into an industrial phenotypic high throughput screening campaign.**

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - CCD-18co cell line
  - CCD-18co cell line maintenance and stimulation
- METHOD DETAILS
  - Single Cell RNA-seq
  - Analysis of CCD-18Co scRNA-Seq data
  - High throughput screening assay
  - Cell Painting high content imaging, feature extraction and high dimensionality data analysis
  - Cell Painting data projection score calculation
  - CCD-18co cell hit confirmation and validation hit picking strategy
  - Visualization of primary screen Cell Painting dataset with t-SNE
  - CXCL10 and cellular feature statistical dependence analysis
  - CRISPR/Cas9, mismatch detection and T7E1 assays
  - Olink Target 96 inflammation assay
  - RNA-seq
  - Pro-fibrotic biomarker detection with Luminex®
  - Immunofluorescence detection for ACTA2 and COL1A1
- QUANTIFICATION AND STATISTICAL ANALYSIS

### SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.chembiol.2023.06.014>.

### ACKNOWLEDGMENTS

We would like to thank our Takeda Development Center of Americas' colleagues Darren Ruane and Christopher Haines for suggesting the pro-fibrotic stimuli to be used for this study, Tony Li for CCD-18co cell culture for RNA-seq experiment, and Narender Gavva for the general supervision of the Early Target Discovery/Core Biology group and administrative support. We would also like to thank Joseph Z. Chen from California Institute for Biomedical Research, a division of Scripps Research for designing the graphic abstract.

Authors from the Broad Institute of Harvard and MIT were funded by a grant from Takeda Development Center Americas for part of this work, as well as by the National Institutes of Health (R35 GM122547 to A.E.C.).

### AUTHOR CONTRIBUTIONS

S.Y., S.R.H., and D.N. conceived and designed the project; S.Y., A.K., M.P., M.M., and S.R.H. contributed to the result interpretation and drafted the manuscript; S.Y. performed assay development and high throughput screen experiments; Y.L. participated in high throughput screen experiments; S.Y., Y.L., and Q.W. performed CRISPR/Cas9 experiment; M.M. designed, performed, and analyzed scRNA-seq experiment; M.P. performed transcriptomic analysis of CCD-18co datasets and public human colon biopsies datasets; I.I. performed PCA on pro-fibrotic stimuli treated CCD-18co Cell Painting data; J.T. performed PCA on CCD-18co RNA-seq data; J.C. performed Cell Painting primary hit picking with linear predictive model; D.S. and J.H. assisted with laboratory automation and compound management; S.S. and A.E.C. provided input on morphological profiling analysis and its result interpretation, and edited the manuscript.

### DECLARATION OF INTERESTS

We declare competing interests. The authors who were affiliated with Takeda Development Center Americas were employees of Takeda Pharmaceuticals during the course of this work, and have real or potential ownership interest in Takeda. AEC serves on the Scientific Advisory Board of, and has ownership interest in, Recursion, a pharmaceutical company using image-based profiling

for drug discovery. Authors from the Broad Institute of Harvard and MIT were funded by a grant from Takeda for part of this work.

### INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: July 24, 2022

Revised: March 11, 2023

Accepted: June 13, 2023

Published: July 11, 2023

### REFERENCES

- Lenti, M.V., and Di Sabatino, A. (2019). Intestinal fibrosis. *Mol. Aspects Med.* **65**, 100–109.
- Pariante, B., Cosnes, J., Danese, S., Sandborn, W.J., Lewin, M., Fletcher, J.G., Chowers, Y., D'Haens, G., Feagan, B.G., Hibi, T., et al. (2011). Development of the Crohn's disease digestive damage score, the Lemann score. *Inflamm. Bowel Dis.* **17**, 1415–1422.
- Rieder, F., and Fiocchi, C. (2009). Intestinal fibrosis in IBD—a dynamic, multifactorial process. *Nat. Rev. Gastroenterol. Hepatol.* **6**, 228–235.
- Kaplan, G.G., and Windsor, J.W. (2021). The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 56–66.
- Abdulla, M., and Chew, T.S. (2021). Molecular targets and the use of biologics in the management of small bowel fibrosis in inflammatory bowel disease. *Curr. Opin. Gastroenterol.* **37**, 275–283.
- Friedrich, M., Pohin, M., and Powrie, F. (2019). Cytokine Networks in the Pathophysiology of Inflammatory Bowel Disease. *Immunity* **50**, 992–1006.
- Hayashi, Y., and Nakase, H. (2022). The Molecular Mechanisms of Intestinal Inflammation and Fibrosis in Crohn's Disease. *Front. Physiol.* **13**, 845078.
- Walton, K.L., Johnson, K.E., and Harrison, C.A. (2017). Targeting TGF-beta Mediated SMAD Signaling for the Prevention of Fibrosis. *Front. Pharmacol.* **8**, 461.
- Kim, M.H., Jung, S.Y., Song, K.H., Park, J.I., Ahn, J., Kim, E.H., Park, J.K., Hwang, S.G., Woo, H.J., and Song, J.Y. (2020). A new FGFR inhibitor disrupts the TGF-beta1-induced fibrotic process. *J. Cell Mol. Med.* **24**, 830–840.
- Yu, S., Ericson, M., Fanjul, A., Erion, D.M., Paraskevopoulou, M., Smith, E.N., Cole, B., Feaver, R., Holub, C., Gavva, N., et al. (2022). Genome-wide CRISPR Screening to Identify Drivers of TGF-beta-Induced Liver Fibrosis in Human Hepatic Stellate Cells. *ACS Chem. Biol.* **17**, 918–929.
- Dewidar, B., Meyer, C., Dooley, S., and Meindl-Beinker, A.N. (2019). TGF-beta in Hepatic Stellate Cell Activation and Liver Fibrogenesis-Updated 2019. *Cells* **8**, 1419.
- Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714–730.e22.
- Wei, K., Nguyen, H.N., and Brenner, M.B. (2021). Fibroblast pathology in inflammatory diseases. *J. Clin. Invest.* **131**, e149538.
- Yamaguchi, S., Kaneko, M., and Narukawa, M. (2021). Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clin. Transl. Sci.* **14**, 1113–1122.
- Bunnage, M.E. (2011). Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* **7**, 335–339.
- Cong, F., Cheung, A.K., and Huang, S.M.A. (2012). Chemical genetics-based target identification in drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **52**, 57–78.
- Jones, L.H., and Bunnage, M.E. (2017). Applications of chemogenomic library screening in drug discovery. *Nat. Rev. Drug Discov.* **16**, 285–296.
- Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D., and Carpenter, A.E. (2021). Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159.
- Hinterleitner, T.A., Saada, J.I., Berschneider, H.M., Powell, D.W., and Valentich, J.D. (1996). IL-1 stimulates intestinal myofibroblast COX gene expression and augments activation of Cl- secretion in T84 cells. *Am. J. Physiol.* **271**, C1262–C1268.
- Groover, M.K., and Richmond, J.M. (2020). Potential therapeutic manipulations of the CXCR3 chemokine axis for the treatment of inflammatory fibrosing diseases. *F1000Res.* **9**, 1197.
- Julian, D.R., Kazakoff, M.A., Patel, A., Jaynes, J., Willis, M.S., and Yates, C.C. (2021). Chemokine-Based Therapeutics for the Treatment of Inflammatory and Fibrotic Convergent Pathways in COVID-19. *Curr. Pathobiol. Rep.* **9**, 93–105.
- Korsunsky, I., Wei, K., Pohin, M., Kim, E.Y., Barone, F., Major, T., Taylor, E., Ravindran, R., Kemble, S., Watts, G.F.M., et al. (2022). Cross-tissue, single-cell stromal atlas identifies shared pathological fibroblast phenotypes in four chronic inflammatory diseases. *Med (N Y)* **3**, 481–518.e14.
- Raghu, G., Martinez, F.J., Brown, K.K., Costabel, U., Cottin, V., Wells, A.U., Lancaster, L., Gibson, K.F., Haddad, T., Agarwal, P., et al. (2015). CC-chemokine ligand 2 inhibition in idiopathic pulmonary fibrosis: a phase 2 trial of carlumab. *Eur. Respir. J.* **46**, 1740–1750.
- Berres, M.L., Trautwein, C., Schmeding, M., Eurich, D., Tacke, F., Bahra, M., Neuhaus, P., Neumann, U.P., and Wasmuth, H.E. (2011). Serum chemokine CXC ligand 10 (CXCL10) predicts fibrosis progression after liver transplantation for hepatitis C infection. *Hepatology* **53**, 596–603.
- Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., and Hopkins, A.L. (2012). Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98.
- Bray, M.A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., and Carpenter, A.E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774.
- Zhang, X.D. (2011). Illustration of SSMD, z score, SSMD\*, z\* score, and t statistic for hit selection in RNAi high-throughput screens. *J. Biomol. Screen* **16**, 775–785.
- Simm, J., Klambauer, G., Arany, A., Steijaert, M., Wegner, J.K., Gustin, E., Chupakhin, V., Chong, Y.T., Vialard, J., Buijsters, P., et al. (2018). Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem. Biol.* **25**, 611–618.e3.
- Amano, H., Matsui, Y., Hatanaka, K., Hosono, K., and Ito, Y. (2021). VEGFR1-tyrosine kinase signaling in pulmonary fibrosis. *Inflamm. Regen.* **41**, 16.
- Yang, L., Zhou, F., Zheng, D., Wang, D., Li, X., Zhao, C., and Huang, X. (2021). FGF/FGFR signaling: From lung development to respiratory diseases. *Cytokine Growth Factor Rev.* **62**, 94–104.
- Liang, D., Chen, H., Zhao, L., Zhang, W., Hu, J., Liu, Z., Zhong, P., Wang, W., Wang, J., and Liang, G. (2018). Inhibition of EGFR attenuates fibrosis and stellate cell activation in diet-induced model of nonalcoholic fatty liver disease. *Biochim. Biophys. Acta, Mol. Basis Dis.* **1864**, 133–142.
- Hostettler, K.E., Zhong, J., Papakonstantinou, E., Karakiulakis, G., Tamm, M., Seidel, P., Sun, Q., Mandal, J., Lardinio, D., Lambers, C., and Roth, M. (2014). Anti-fibrotic effects of nintedanib in lung fibroblasts derived from patients with idiopathic pulmonary fibrosis. *Respir. Res.* **15**, 157.
- Rebeyrol, C., Saint-Criq, V., Guillot, L., Riffault, L., Corvol, H., Chadelat, K., Ray, D.W., Clement, A., Tabary, O., and Le Rouzic, P. (2012). Glucocorticoids reduce inflammation in cystic fibrosis bronchial epithelial cells. *Cell. Signal.* **24**, 1093–1099.
- Vaglio, A., Palmisano, A., Alberici, F., Maggiore, U., Ferretti, S., Cobelli, R., Ferrozzi, F., Corradi, D., Salvarani, C., and Buzio, C. (2011). Prednisone versus tamoxifen in patients with idiopathic retroperitoneal fibrosis: an open-label randomised controlled trial. *Lancet* **378**, 338–346.
- van Bommel, E.F.H., Siemes, C., Hak, L.E., van der Veer, S.J., and Hendriks, T.R. (2007). Long-term renal and patient outcome in idiopathic



- retroperitoneal fibrosis treated with prednisone. *Am. J. Kidney Dis.* **49**, 615–625.
36. Giebeler, A., Boekschoten, M.V., Klein, C., Borowiak, M., Birchmeier, C., Gassler, N., Wasmuth, H.E., Müller, M., Trautwein, C., and Streetz, K.L. (2009). c-Met confers protection against chronic liver tissue damage and fibrosis progression after bile duct ligation in mice. *Gastroenterology* **137**, 297–308.e1–4.
  37. Kawaguchi, Y., Harigai, M., Hara, M., Fukasawa, C., Takagi, K., Tanaka, M., Tanaka, E., Nishimagi, E., and Kamatani, N. (2002). Expression of hepatocyte growth factor and its receptor (c-met) in skin fibroblasts from patients with systemic sclerosis. *J. Rheumatol.* **29**, 1877–1883.
  38. Zbodakova, O., Chalupsky, K., Sarnova, L., Kasperek, P., Jirouskova, M., Gregor, M., and Sedlacek, R. (2021). ADAM10 and ADAM17 regulate EGFR, c-Met and TNF RI signalling in liver regeneration and fibrosis. *Sci. Rep.* **11**, 11414.
  39. Ji, D., Zhao, Q., Qin, Y., Tong, H., Wang, Q., Yu, M., Mao, C., Lu, T., Qiu, J., and Jiang, C. (2021). Germacrone improves liver fibrosis by regulating the PI3K/AKT/mTOR signalling pathway. *Cell Biol. Int.* **45**, 1866–1875.
  40. Lukey, P.T., Harrison, S.A., Yang, S., Man, Y., Holman, B.F., Rashidnasab, A., Azzopardi, G., Grayer, M., Simpson, J.K., Bareille, P., et al. (2019). A randomised, placebo-controlled study of omipalisib (PI3K/mTOR) in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **53**, 1801992.
  41. Madala, S.K., Edukulla, R., Phatak, M., Schmidt, S., Davidson, C., Acciani, T.H., Korfhagen, T.R., Medvedovic, M., Lecras, T.D., Wagner, K., and Hardie, W.D. (2014). Dual targeting of MEK and PI3K pathways attenuates established and progressive pulmonary fibrosis. *PLoS One* **9**, e86536.
  42. Qin, W., Cao, L., and Massey, I.Y. (2021). Role of PI3K/Akt signaling pathway in cardiac fibrosis. *Mol. Cell. Biochem.* **476**, 4045–4059.
  43. Wang, J., Hu, K., Cai, X., Yang, B., He, Q., Wang, J., and Weng, Q. (2022). Targeting PI3K/AKT signaling for treatment of idiopathic pulmonary fibrosis. *Acta Pharm. Sin. B* **12**, 18–32.
  44. Espindola, M.S., Habiél, D.M., Narayanan, R., Jones, I., Coelho, A.L., Murray, L.A., Jiang, D., Noble, P.W., and Hogaboam, C.M. (2018). Targeting of TAM Receptors Ameliorates Fibrotic Mechanisms in Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **197**, 1443–1456.
  45. Mukherjee, S.K., Wilhelm, A., and Antoniadis, C.G. (2016). TAM receptor tyrosine kinase function and the immunopathology of liver disease. *Am. J. Physiol. Gastrointest. Liver Physiol.* **310**, G899–G905.
  46. Hu, M., Che, P., Han, X., Cai, G.Q., Liu, G., Antony, V., Luckhardt, T., Siegal, G.P., Zhou, Y., Liu, R.M., et al. (2014). Therapeutic targeting of SRC kinase in myofibroblast differentiation and pulmonary fibrosis. *J. Pharmacol. Exp. Ther.* **351**, 87–95.
  47. Li, H., Zhao, C., Tian, Y., Lu, J., Zhang, G., Liang, S., Chen, D., Liu, X., Kuang, W., and Zhu, M. (2020). Src family kinases and pulmonary fibrosis: A review. *Biomed. Pharmacother.* **127**, 110183.
  48. Lu, Y.Y., Zhao, X.K., Yu, L., Qi, F., Zhai, B., Gao, C.Q., and Ding, Q. (2017). Interaction of Src and Alpha-V Integrin Regulates Fibroblast Migration and Modulates Lung Fibrosis in A Preclinical Model of Lung Fibrosis. *Sci. Rep.* **7**, 46357.
  49. Chen, J., Chen, J.K., Nagai, K., Plieth, D., Tan, M., Lee, T.C., Threadgill, D.W., Neilson, E.G., and Harris, R.C. (2012). EGFR signaling promotes TGFbeta-dependent renal fibrosis. *J. Am. Soc. Nephrol.* **23**, 215–224.
  50. Vallath, S., Hynds, R.E., Sucony, L., Janes, S.M., and Giangreco, A. (2014). Targeting EGFR signalling in chronic lung disease: therapeutic challenges and opportunities. *Eur. Respir. J.* **44**, 513–522.
  51. Kim, S.K., Jung, S.M., Park, K.S., and Kim, K.J. (2021). Integrative analysis of lung molecular signatures reveals key drivers of idiopathic pulmonary fibrosis. *BMC Pulm. Med.* **21**, 404.
  52. Liu, Y., Li, J., Liao, L., Huang, H., Fan, S., Fu, R., Huang, J., Shi, C., Yu, L., Chen, K.X., et al. (2021). Cyclin-dependent kinase inhibitor roscovitine attenuates liver inflammation and fibrosis by influencing initiating steps of liver injury. *Clin. Sci.* **135**, 925–941.
  53. Madala, S.K., Schmidt, S., Davidson, C., Ikegami, M., Wert, S., and Hardie, W.D. (2012). MEK-ERK pathway modulation ameliorates pulmonary fibrosis associated with epidermal growth factor receptor activation. *Am. J. Respir. Cell Mol. Biol.* **46**, 380–388.
  54. Terashima, H., Aonuma, M., Tsuchida, H., Sugimoto, K., Yokoyama, M., and Kato, M. (2019). Attenuation of pulmonary fibrosis in type I collagen-targeted reporter mice with ALK-5 inhibitors. *Pulm. Pharmacol. Ther.* **54**, 31–38.
  55. Beyer, C., and Distler, J.H.W. (2013). Tyrosine kinase signaling in fibrotic disorders: Translation of basic research to human disease. *Biochim. Biophys. Acta* **1832**, 897–904.
  56. Mansuroglu, T., Ramadori, P., Dudás, J., Malik, I., Hammerich, K., Füzési, L., and Ramadori, G. (2009). Expression of stem cell factor and its receptor c-Kit during the development of intrahepatic cholangiocarcinoma. *Lab. Invest.* **89**, 562–574.
  57. Huang, L.S., Fu, P., Patel, P., Harijith, A., Sun, T., Zhao, Y., Garcia, J.G.N., Chun, J., and Natarajan, V. (2013). Lysophosphatidic acid receptor-2 deficiency confers protection against bleomycin-induced lung injury and fibrosis in mice. *Am. J. Respir. Cell Mol. Biol.* **49**, 912–922.
  58. Sakai, N., Chun, J., Duffield, J.S., Lagares, D., Wada, T., Luster, A.D., and Tager, A.M. (2017). Lysophosphatidic acid signaling through its receptor initiates profibrotic epithelial cell fibroblast communication mediated by epithelial cell derived connective tissue growth factor. *Kidney Int.* **91**, 628–641.
  59. Goodwin, J., Choi, H., Hsieh, M.H., Neugent, M.L., Ahn, J.M., Hayenga, H.N., Singh, P.K., Shackelford, D.B., Lee, I.K., Shulaev, V., et al. (2018). Targeting Hypoxia-Inducible Factor-1alpha/Pyruvate Dehydrogenase Kinase 1 Axis by Dichloroacetate Suppresses Bleomycin-induced Pulmonary Fibrosis. *Am. J. Respir. Cell Mol. Biol.* **58**, 216–231.
  60. Hamanaka, R.B., Nigdeliöglu, R., Meliton, A.Y., Tian, Y., Witt, L.J., O’Leary, E., Sun, K.A., Woods, P.S., Wu, D., Ansbro, B., et al. (2018). Inhibition of Phosphoglycerate Dehydrogenase Attenuates Bleomycin-induced Pulmonary Fibrosis. *Am. J. Respir. Cell Mol. Biol.* **58**, 585–593.
  61. Buhl, E.M., Djurdjaj, S., Klinkhammer, B.M., Ermert, K., Puelles, V.G., Lindenmeyer, M.T., Cohen, C.D., He, C., Borkham-Kamphorst, E., Weiskirchen, R., et al. (2020). Dysregulated mesenchymal PDGFR-beta drives kidney fibrosis. *EMBO Mol. Med.* **12**, e11021.
  62. Xie, Y., Su, N., Yang, J., Tan, Q., Huang, S., Jin, M., Ni, Z., Zhang, B., Zhang, D., Luo, F., et al. (2020). FGF/FGFR signaling in health and disease. *Signal Transduct. Target. Ther.* **5**, 181.
  63. Tanjore, H., Blackwell, T.S., and Lawson, W.E. (2012). Emerging evidence for endoplasmic reticulum stress in the pathogenesis of idiopathic pulmonary fibrosis. *Am. J. Physiol. Lung Cell Mol. Physiol.* **302**, L721–L729.
  64. Kropski, J.A., and Blackwell, T.S. (2018). Endoplasmic reticulum stress in the pathogenesis of fibrotic disease. *J. Clin. Invest.* **128**, 64–73.
  65. Santacroce, G., Lenti, M.V., and Di Sabatino, A. (2022). Therapeutic Targeting of Intestinal Fibrosis in Crohn’s Disease. *Cells* **11**, 429.
  66. Flaherty, K.R., Wells, A.U., Cottin, V., Devaraj, A., Walsh, S.L.F., Inoue, Y., Richeldi, L., Kolb, M., Tetzlaff, K., Stowasser, S., et al. (2019). Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *N. Engl. J. Med.* **381**, 1718–1727.
  67. Kurita, Y., Araya, J., Minagawa, S., Hara, H., Ichikawa, A., Saito, N., Kadota, T., Tsubouchi, K., Sato, N., Yoshida, M., et al. (2017). Pirfenidone inhibits myofibroblast differentiation and lung fibrosis development during insufficient mitophagy. *Respir. Res.* **18**, 114.
  68. Kalderén, C., Stadler, C., Forsgren, M., Kvastad, L., Johansson, E., Sydow-Bäckman, M., and Svensson Gélius, S. (2014). CCL2 mediates anti-fibrotic effects in human fibroblasts independently of CCR2. *Int. Immunopharmacol.* **20**, 66–73.
  69. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95.
  70. Waskom, M. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021.

71. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., and Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* *12*, 2825–2830.
72. McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, 445, pp. 51–56.
73. Wali, G., Berkovsky, S., Whiten, D.R., Mackay-Sim, A., and Sue, C.M. (2021). Single cell morphology distinguishes genotype and drug effect in Hereditary Spastic Paraplegia. *Sci. Rep.* *11*, 16635.
74. Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to Use t-SNE Effectively. *Distill* *1*.



# Resolve the complexity of disease through morphology

Achieve unparalleled new insight into disease with morphological profiling and AI. Uncover and isolate rare and unique cell populations, identify new drug targets and drug resistance mechanisms, or incorporate morphology into patient stratification models. VisionSort elevates biomarker discovery campaigns to a new level with flexible, easy-to-use, and user-controlled AI algorithms embedded directly in the instrument. Discover more with unbiased morphometric characterization.



## VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT

[THINKCYTE.COM](http://THINKCYTE.COM)



# See Your Cells in a Whole New Light

VisionSort brings together fundamental advances in optics, microfluidics, and artificial intelligence (AI) to deliver morphological profiling and label-free cell sorting on top of the capabilities you have come to expect from traditional fluorescence-only cytometers. Lose nothing, gain everything with VisionSort.



## VisionSort™

Dual Mode Cell Sorter  
Fluorescence & Label-Free  
Morphotypic Cell Sorting

LEARN MORE AT

[THINKCYTE.COM](http://THINKCYTE.COM)