Trends in
Pharmacological Sciences

CellPress
OPEN ACCESS

Opinion

AI-powered therapeutic target discovery

Frank W. Pun,[1] Ivan V. Ozerov,[1] and Alex Zhavoronkov [1,2,3,*]

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.

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.

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%[i]. 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[ii,iii]. 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.

[1]Insilico Medicine Hong Kong Ltd., Hong Kong Science and Technology Park, New Territories, Hong Kong
[2]Insilico Medicine MENA, 6F IRENA Building, Abu Dhabi, United Arab Emirates
[3]Buck Institute for Research on Aging, Novato, CA, USA

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

https://doi.org/10.1016/j.tips.2023.06.010

© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

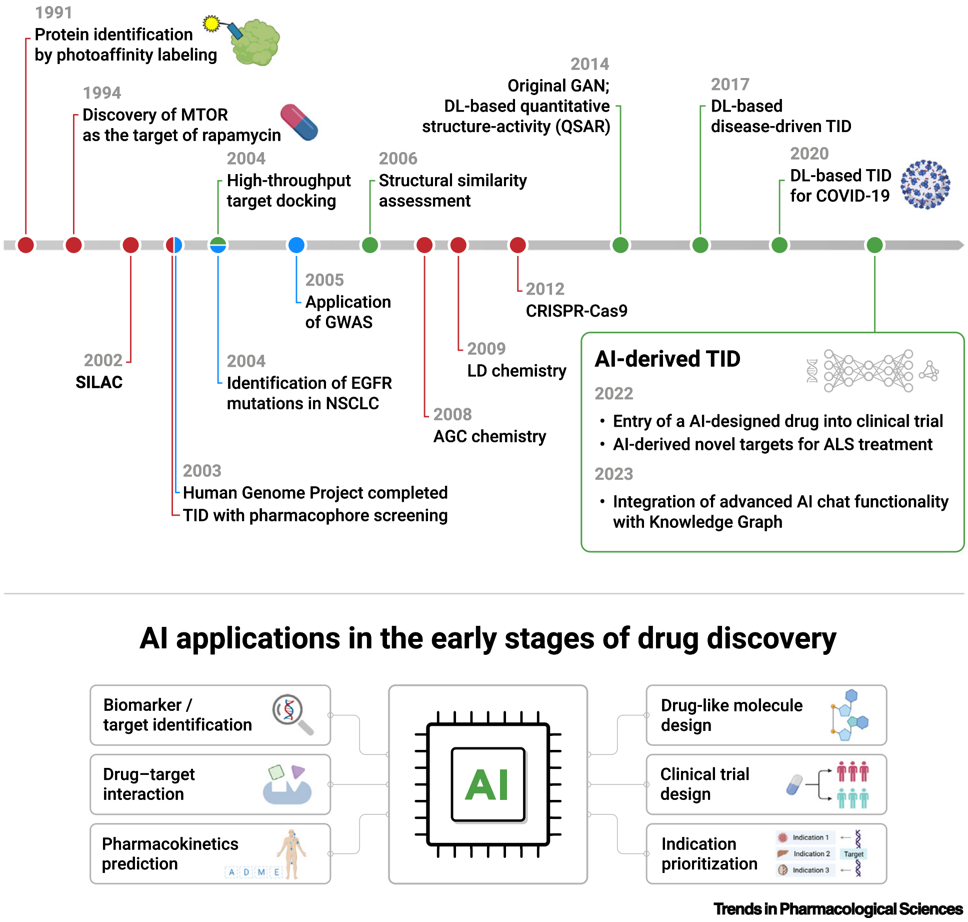

1991
Protein identification by photoaffinity labeling
1994
Discovery of MTOR as the target of rapamycin
2004
High-throughput target docking
2006
Structural similarity assessment
2014
Original GAN; DL-based quantitative structure-activity (QSAR)
2017
DL-based disease-driven TID
2020
DL-based TID for COVID-19
2005
Application of GWAS
2012
CRISPR-Cas9
2002
SILAC
2004
Identification of EGFR mutations in NSCLC
2009
LD chemistry
2008
AGC chemistry
2003
Human Genome Project completed
TID with pharmacophore screening
AI-derived TID
2022
• Entry of a AI-designed drug into clinical trial
• AI-derived novel targets for ALS treatment
2023
• Integration of advanced AI chat functionality with Knowledge Graph
AI applications in the early stages of drug discovery
Biomarker / target identification
Drug–target interaction
Pharmacokinetics prediction
A D M E
AI
Drug-like molecule design
Clinical trial design
Indication prioritization
Indication 1
Indication 2
Indication 3
Target
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.

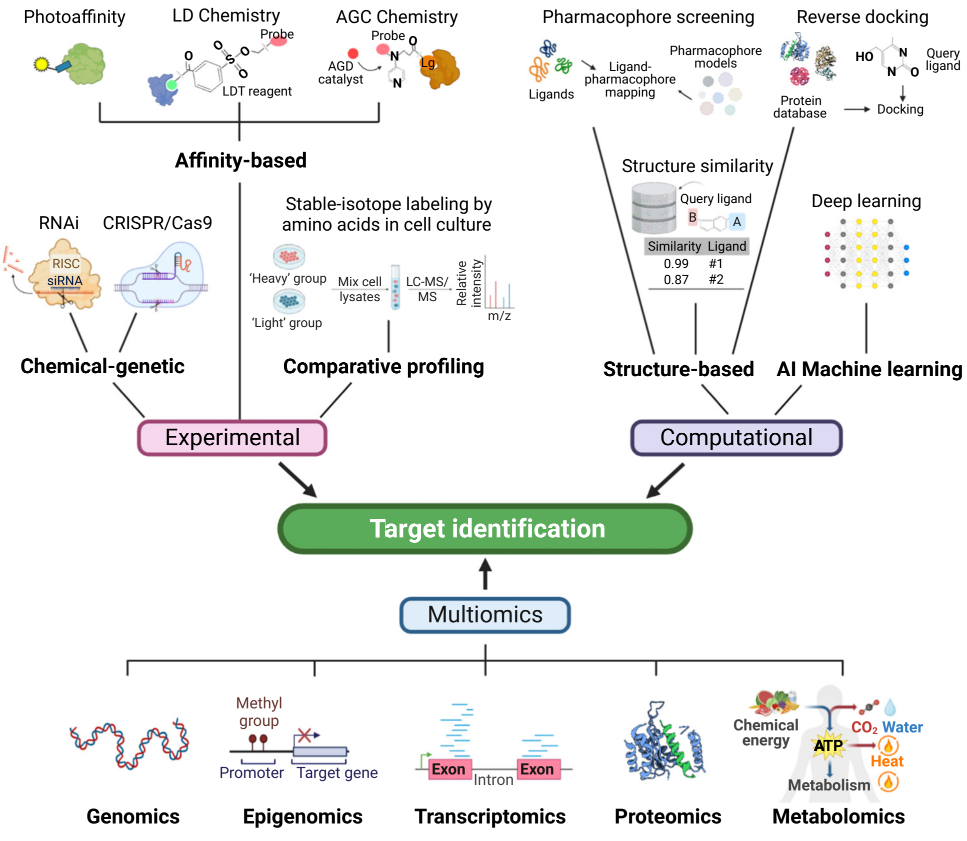
Indication prioritization: the process of prioritizing the potential indications of a drug based on the expected relevancy of the drug and a specific indication using AI.

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.



Trends in Pharmacological Sciences

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

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.

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[a]	Compound	Development status	Trial number[b]
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[vi]
	5-HT1A/2A	Alzheimer's disease psychosis	DSP-0038	Phase 1	Undisclosed[vii]
	PKC-θ	Inflammatory diseases	EXS4318	Phase 1/2	Undisclosed[viii]
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
	PI3Kα	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

[a] Indications retrieved from the company pipeline.
[b] 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[iv], 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

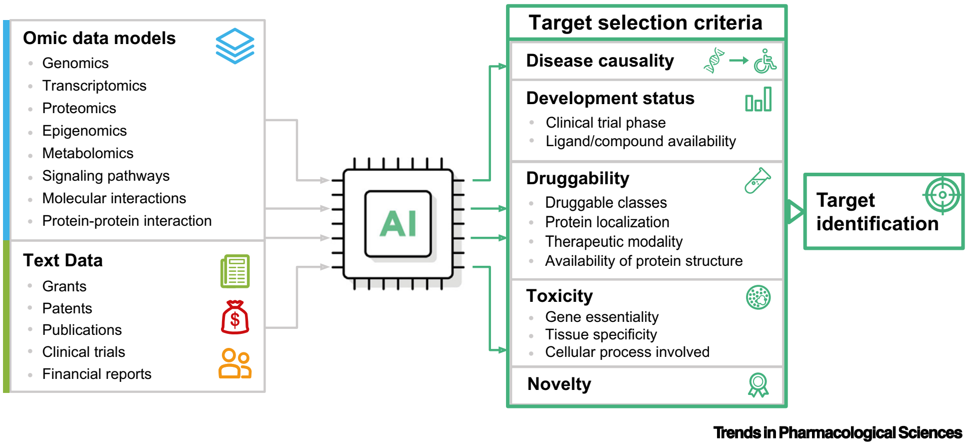


Figure 3. AI prioritizes targets for specific indications by using multi-models that utilize a diverse range of publicly available omic and text data. Omic data encompass genomics, transcriptomics, proteomics, epigenomics, and metabolomics. These data provide information about altered signaling pathways, 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[ix]. 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-γt 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[v]. 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

[i] https://ftloscience.com/process-costs-drug-development/

[ii] www.fiercebiotech.com/medtech/breaking-big-pharma-s-ai-barrier-insilico-medicine-uncovers-novel-target-new-drug-for

[iii] www.nature.com/articles/d43747-021-00045-7

[iv] www.eurekalert.org/news-releases/982543

[v] www.prnewswire.com/news-releases/insilico-medicine-announces-novel-3cl-protease-inhibitor-preclinical-candidate-for-covid-19-treatment-301553766.html

[vi] www.exscientia.ai/dsp-1181

[vii] 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

[viii] 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

[ix] 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

1. Hinkson, I.V. *et al.* (2020) Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. *Front. Pharmacol.* 11, 770
2. 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
3. Kana, O. and Brylinski, M. (2019) Elucidating the druggability of the human proteome with eFindSite. *J. Comput. Aided Mol. Des.* 33, 509–519
4. Finan, C. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9, eaag1166
5. Sun, D. *et al.* (2022) Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* 12, 3049–3062
6. Shiraiwa, K. *et al.* (2020) Chemical tools for endogenous protein labeling and profiling. *Cell Chem. Biol.* 27, 970–985
7. van der Zouwen, A.J. and Witte, M.D. (2021) Modular approaches to synthesize activity- and affinity-based chemical probes. *Front. Chem.* 9, 644811
8. 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
9. Jin, J. *et al.* (2023) SIRT3-dependent delactylation of cyclin E2 prevents hepatocellular carcinoma growth. *EMBO Rep.* 24, e56052
10. Li, X. *et al.* (2023) Deacetylation induced nuclear condensation of HP1gamma promotes multiple myeloma drug resistance. *Nat. Commun.* 14, 1290
11. Wang, Y. *et al.* (2022) DUT enhances drug resistance to proteasome inhibitors via promoting mitochondrial function in multiple myeloma. *Carcinogenesis* 43, 1030–1038
12. 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
13. Kortum, B. *et al.* (2022) Combinatorial treatment with statins and niclosamide prevents CRC dissemination by unhinging the MACC1–beta-catenin–S100A4 axis of metastasis. *Oncogene* 41, 4446–4458
14. 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
15. 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
16. 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
17. Ramkumar, P. *et al.* (2020) CRISPR-based screens uncover determinants of immunotherapy response in multiple myeloma. *Blood Adv.* 4, 2899–2911
18. Chakraborty, S. *et al.* (2018) Onco-Multi-OMICS approach: a new frontier in cancer research. *Biomed. Res. Int.* 2018, 9836256
19. Nurk, S. *et al.* (2022) The complete sequence of a human genome. *Science* 376, 44–53
20. 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
21. 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
22. 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
23. Deelen, J. *et al.* (2019) Publisher correction: a meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* 10, 3669
24. 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
25. Abell, N.S. *et al.* (2022) Multiple causal variants underlie genetic associations in humans. *Science* 375, 1247–1254
26. Assum, I. *et al.* (2022) Tissue-specific multi-omics analysis of atrial fibrillation. *Nat. Commun.* 13, 441
27. Suhre, K. *et al.* (2017) Erratum: connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* 8, 14357
28. 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
29. 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
30. Na, F. *et al.* (2022) KMT2C deficiency promotes small cell lung cancer metastasis through DNMT3A-mediated epigenetic reprogramming. *Nat. Can.* 3, 753–767
31. Gulfidan, G. *et al.* (2022) Systems biomarkers for papillary thyroid cancer prognosis and treatment through multi-omics networks. *Arch. Biochem. Biophys.* 715, 109085
32. Lu, J. *et al.* (2021) Multi-omics analysis of fatty acid metabolism in thyroid carcinoma. *Front. Oncol.* 11, 737127
33. 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
34. Pinero, J. *et al.* (2018) Network, transcriptomic and genomic features differentiate genes relevant for drug response. *Front. Genet.* 9, 412
35. Wolber, G. *et al.* (2008) Molecule–pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* 13, 23–29
36. Lee, A. *et al.* (2016) Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov.* 11, 707–715
37. 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
38. Lo, Y.C. *et al.* (2016) 3D chemical similarity networks for structure-based target prediction and scaffold hopping. *ACS Chem. Biol.* 11, 2244–2253
39. Vamathevan, J. *et al.* (2019) Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477
40. 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
41. Zhavoronkov, A. *et al.* (2019) Deep biomarkers of aging and longevity: from research to applications. *Aging (Albany NY)* 11, 10771–10780
42. Muslu, O. *et al.* (2022) GuiltyTargets: prioritization of novel therapeutic targets with network representation learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 491–500
43. 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
44. 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 oversampling 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. Buphamalai, 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