AI Approaches for the Discovery and Validation of Drug Targets

A. Wenteler [1, 2, 3 *], C.P. Cabrera [1, 2, 4], W. Wei [3], V. Neduva [3], M.R. Barnes [1, 2, 4, 5]

[1] Digital Environment Research Institute, Queen Mary University of London, London, United Kingdom
[2] Centre for Translational Bioinformatics, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom
[3] MSD Discovery Centre, London, United Kingdom
[4] NIHR Barts Cardiovascular Biomedical Research Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom
[5] The Alan Turing Institute, London, United Kingdom

[*] Corresponding author: a.wenteler@qmul.ac.uk

Abstract

Artificial intelligence (AI) holds immense promise for accelerating and improving all aspects of drug discovery, not least target discovery and validation. By integrating a diverse range of biological data modalities, AI enables the accurate prediction of drug target properties, ultimately illuminating biological mechanisms of disease and guiding drug discovery strategies. Despite the indisputable potential of AI in drug target discovery, there are many challenges and obstacles yet to be overcome, including dealing with data biases, model interpretability and generalisability, and the validation of predicted drug targets to name a few. By exploring recent advancements in AI, this review showcases current applications of AI for drug target discovery and offers perspectives on the future of AI for the discovery and validation of drug targets, paving the way for the generation of novel and safer pharmaceuticals.

Keywords: Drug discovery, Drug targets, Artificial intelligence, Machine learning, Multiomics

This peer-reviewed article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI.
10.1017/pcm.2024.4

Background

Historically, drug target discovery and validation has been a laborious and somewhat haphazard process, heavily reliant on industry standard laboratory models and analysis procedures (Drews 2000; Huang *et al.* 2004; Materi and Wishart 2007). Most drug discovery to date has taken a phenotype-first approach focusing on evaluation of the therapeutic potential of compounds in phenotypic assays, without necessarily knowing the exact target or mechanism of action (Moffat *et al.* 2017). This approach relies largely on serendipity, where complex compound libraries, including phytochemicals, biochemicals and other organic chemistry, are identified for therapeutic use by chance. Naturally, pharma companies initially sought to improve their odds by increasing the size and complexity of their compound libraries, and by the mid-2000s most major pharmaceutical companies had compound libraries in the range of 1-2 million small molecule entities (SMEs) (Hann and Oprea 2004). However, the unsustainability of this chemistry arms race has spurred a shift towards a target-first strategy, which signified a pivotal moment in pharmacological research, emphasising the importance of thorough understanding and validation of a biological target before initiation of the drug design process. This paradigm shift marked a transition from empirical, trial-and-error methods to a more rational and systematic approach, greatly enhancing the efficiency and effectiveness of drug discovery. Ironically, although the target-first approach was designed to reduce the complexity of drug discovery, it may have had the opposite effect, simply highlighting the challenges of true target validation, leading to over a decade of increased failure in drug discovery stemming from poorly validated targets (Paul *et al.* 2010; Scannell *et al.* 2012). With an increasing repertoire of biomolecular assays to probe mechanism, such as CRISPR-Cas9, so-called target deconvolution studies have been conducted. These studies connect phenotypic to target-first approaches by attempting to elucidate the mechanism of action of the target upon which a drug acted retrospectively. This strategy enriches the phenotype-centric drug discovery paradigm with mechanistic understanding of the observed therapeutic effect and set the groundwork for integration of phenotype-first and target-first approaches (Terstappen *et al.* 2007).

In this review, we define drug targets as biomolecules—primarily proteins, but also DNA, RNA, or other biomolecular species—that a therapeutic compound can bind to or modulate. The pool of existing drug targets is limited, and assessments of the druggable genome, which refers to those genes susceptible to modulation by small

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

molecules, fluctuate. The latest estimate places this number at 4,479 potential targets, accounting for approximately 22% of protein-coding genes (Finan *et al.* 2017). According to records of the Human Protein Atlas (HPA), there are approximately 863 FDA approved drug targets (Paananen and Fortino 2020), over 50% of these targets are represented by just four protein families - ion channels, nuclear receptors, kinases, and G-protein coupled receptors (Bakheet and Doig 2009; Santos *et al.* 2017). When it comes to finding novel, efficacious, and safer drug targets, as a general guideline, targets should have a role in disease, limited role in normal physiology, particularly in critical tissues such as the heart, and ideally should be druggable with small molecules, although biologic drugs and gene targeted therapies make almost all targets therapeutically tractable. Furthermore, while a laboratory resolved 3D protein structure was a prior requirement for drug design, with the advent of protein structure prediction models, further accelerated by AI approaches (Baek *et al.* 2021; Jumper *et al.* 2021; Lin *et al.* 2023b), high quality 3D structures of a wide range of potential drug targets are generally available. This enables a broader application of *in silico* structure-based drug design. Another desirable property for a drug target is having multiple binding pockets. By having multiple potential binding pockets, different conformations of the protein in various functional states can be targeted. It also provides opportunities for identifying allosteric inhibitors rather than only targeting the active site. Allosteric sites may offer better selectivity and provide safety benefits (Abdel-Magid 2015; May *et al.* 2007; Wagner *et al.* 2016b). Lastly, by understanding the associated pathways of the target, we gain insight into the processes the target is involved in and thus, what other biological processes could potentially be affected. This can help the assessment of potential off-target effects.

Despite the great progress in drug discovery, the process is still burdened by high costs, long timelines, and extraordinarily high attrition rates in clinical trials, attributed to limited efficacy, safety concerns, off-target effects, or sometimes purely economic reasons (DiMasi *et al.* 2016; Minikel *et al.* 2024) Collectively, against this backdrop of failure, the need for transformative solutions for drug discovery becomes clear. Especially when we consider our incomplete understanding of target mechanism and the vast chemical space of compounds that can interact with that target.

The role of AI in drug discovery

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Ideally, we would develop a comprehensive mathematical framework to systematically navigate the vast search spaces and intricate interactions inherent to drug discovery. However, realising such a framework has proven to be an immensely challenging endeavour with limited success so far. In contrast, methods using artificial intelligence (AI) are particularly well-suited for modelling the complexities and nuances of drug discovery. When employing AI, we essentially shift our approach: rather than relying on explicit mathematical descriptions of the underlying biology, chemistry, and physics, we leverage AI models to learn and infer patterns directly from data. While adopting data-driven machine learning techniques holds great promise for enhancing drug discovery pipelines, there are also certain trade-offs, such as a lack of transparency in the models and obscured understanding of causality.

AI has potential to accelerate drug discovery by improving the identification of drug candidates and enhancing our understanding of their mechanisms. The increasing volume of diverse biological and chemical data, including genomics, proteomics, metabolomics, electronic health records, and biomedical literature, combined with high-throughput experiments, greatly enhances AI's ability to extract and interpret insights. Notably, recent studies have highlighted the importance of including genetic and genomic data in drug target discovery pipelines (Razuvayevskaya *et al.* 2023). One estimate quantifying the impact genetic evidence has on success of clinical trials, estimated the odds of advancing to a later stage of clinical trials to be 80% higher when genetic evidence for a target is present (Minikel *et al.* 2024). Furthermore, AI can be used to develop *in silico* methods to predict and simulate biological and chemical spaces. Examples of such approaches are cellular and genetic perturbation modelling (Bunne *et al.* 2023; Prasad *et al.* 2022), gene expression prediction (Avsec *et al.* 2021; Kelley *et al.* 2018; Linder *et al.* 2023), variant effect prediction (Brandes *et al.* 2022; Cheng *et al.* 2023; Frazer *et al.* 2021; Lin *et al.* 2023a), protein structure prediction (Baek *et al.* 2021; Jumper *et al.* 2021; Lin *et al.* 2023b), drug-target interaction prediction (Chen *et al.* 2016; Huang *et al.* 2021; Wen *et al.* 2017), and molecular docking simulations for drug design (Corso *et al.* 2023; Gentile *et al.* 2020).

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

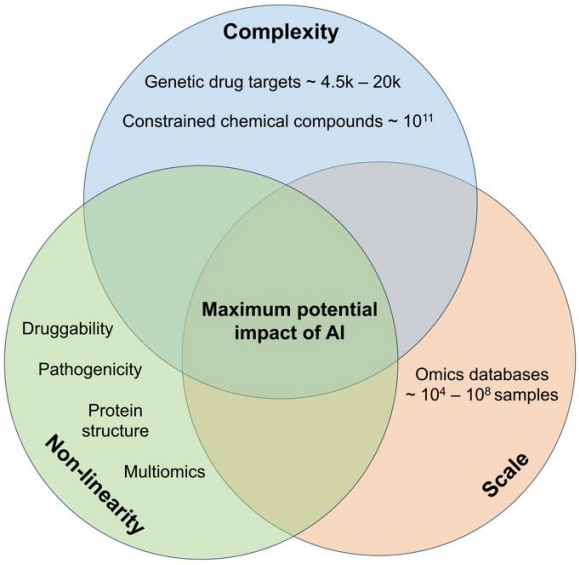

Figure 1: Venn diagram of guiding criteria for the maximum impact of AI in relation to drug discovery. We have made the connection to drug target discovery in the respective sets. The intersection of all sets is where the sweet spot for using AI lies.

When it comes to determining the applicability of AI, we can refer to some guiding principles (Figure 1) that can help us to establish whether introducing AI to solve our problem is sensible. We argue that drug target discovery problems lie at the intersection of all these principles, making them amenable to be solved with AI.

Firstly, the problem at hand must have sufficient scale. Building a successful AI model is reliant on having examples to learn from. While unsupervised approaches can be powerful, the potential of AI predominantly resides in the ability to uncover generalisable patterns within training data through a supervised or a self-supervised framework. A part of this scale is the quality of the data. The dataset should not just be large, but it should also be of
high quality or be processed such that it is of high quality. High quality data implies that the model can learn meaningful signals from the patterns and relationships contained within the data. Some concrete examples of factors potentially decreasing data quality are noise, class imbalances, population bias, and missing data.

Secondly, the complexity of the problem should be appropriate to fully leverage the power of AI models. At the lower bound of the complexity spectrum, the problem could be insufficiently complex, making it likely that an overparameterized AI model is developed that performs seemingly well, but does not generalise. This phenomenon

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

is referred to as overfitting in AI literature. Note that overfitting is not limited to this scenario and can also occur in poorly designed AI models where the problem itself is not necessarily insufficiently complex. At the other end of the complexity spectrum, a problem could be intractable. Take the entire chemical space of $\sim 10^{60}$ compounds for example (Reymond 2015), this immense search space is simply too large for any computational method to fully explore. However, we can make this task more manageable by focusing on a smaller, more relevant subset of compounds. One effective approach to achieve this is by using generative AI models. These models are trained by adding random variations to existing, known data and then attempting to reconstruct the original input from this altered data. Through this process, the model learns the patterns and distributions inherent in the data which can be used to construct outputs based on these patterns.

In the context of drug discovery, this technique can be applied to known chemical structures. This is the basis of Generative Molecular Design (GMD), where AI models are used to generate potentially viable chemical compounds by learning from existing chemical structures (Thomas *et al.* 2023). This approach helps streamline the search for new drug candidates by focusing on the most promising areas of the vast chemical space, in this case up to $\sim 10^{11}$ compounds (Ruddigkeit *et al.* 2012), constraining the search space and thus making the problem computationally tractable. For AI methods to thrive, a balance must be struck as it pertains to the complexity of the problem. We argue that drug discovery, including drug target discovery, satisfies the complexity criterion. Target discovery is often constrained to parameterisations of the genome, or the druggable genome. These are about 20,000 and 4,000 genes in size respectively, which is a tractable search space. As for the chemistry of compounds binding to the target, we can narrow down the search space to effectively design novel compounds.

Lastly, the input features for the problem should be non-linearly related to the target variable. Most biological phenomena are highly non-linear, so it is rare to encounter a biological problem where input and output are linearly related. This also becomes apparent from examining the AI models that underpin some seminal breakthroughs in the context of biology, such as CellOT for gene perturbation prediction (Bunne *et al.* 2023), ESMFold and AlphaFold for protein structure prediction (Jumper *et al.* 2021; Lin *et al.* 2023b), and EVE and AlphaMissense for missense variant pathogenicity

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

prediction (Cheng *et al.* 2023; Frazer *et al.* 2021). To model the non-linearity inherent to these problems, non-linear activation functions are one of the key elements allowing AI models to effectively capture the highly complex relationships within the underlying distributions they attempt to model. Since many biological phenomena exhibit strong non-linearity, it makes sense to express and solve these problems in the language of AI.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

AI Methods and Data Modalities in Drug Target Discovery

One leading reason for the convergence between AI and drug discovery is the diverse range of data types that are being used in drug discovery. The data can be presented in various forms, such as tabular, text, sequences, graphs, and images, each offering a distinct perspective into the biology underlying disease and potential cures. In Table 1, we summarise the different modalities, their use-cases, and some open-access data sources. In the following paragraphs, we briefly discuss each data modality, and how it generally is used in drug target discovery.

One of the most common methods for presenting data related to drug target discovery is through structured tables. Typically, these tabular data structures will contain information describing genes or variants, e.g., allele frequency, mutation type, and conservation scores across species. There are different resources and consortia that aggregate and characterise genomic data in tabular form, such as UK Biobank (Sudlow *et al.* 2015), Genes & Health (Finer *et al.* 2020), and Open Targets (Ochoa *et al.* 2021). Traditional machine learning (ML) methods, e.g. XGBoost (Chen and Guestrin 2016), Linear Regression, Logistic Regression (Pedregosa *et al.* 2011), as well as deep neural networks (LeCun *et al.* 2015), have been developed and tailored to tabular datasets. Therefore, these models have a track record of delivering outstanding performance when working with tabular data.

Textual data, comprising scientific literature, research articles, patents, clinical trial reports, medical textbooks, chemical databases, and electronic health records, represents a valuable resource for drug discovery. The unstructured information in textual documents can provide us with critical insights related to potential drug targets, novel or repurposed drug candidates, and adverse events amongst others. Textual data is typically best analysed using Natural Language Processing (NLP) methods. Recently, Large Language Models (LLMs) have surfaced as the state-of-the-art model type to analyse textual data. LLMs are deep neural networks that combine many different layer types, such as embedding layers, attention layers and linear layers that coalesce to learn semantic information from textual input. Typically, LLMs are pre-trained using self-supervised approaches where a large corpus of text gets tokenised, i.e., it gets mapped to numerical vectors representing the words. This corpus is masked at random, and consequently tasked with predicting the next tokens (Devlin *et al.* 2019; Radford *et al.* 2018). For task-specific objectives, the pre-trained

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

model can be trained further on data related to the task of interest, e.g. information retrieval or translation. (Microsoft Research AI4Science and Microsoft Azure Quantum 2023; Singhal *et al.* 2023a, 2023b)

Data that can be represented sequentially are fundamental to biology. Such sequences often correspond to biological or chemical structure. Some of these data are genomic data, transcriptomic data, protein sequences, and drug compound libraries in the form of SMILES or SELFIES strings. Previously, we introduced language models within the context of natural language. Yet, their versatility transcends the domain of language. Language models also prove adept at understanding biological languages, e.g., decoding semantic meaning from DNA via nucleotide sequences, and unravelling structural or functional information for proteins through the interpretation of amino acid sequences. To model and use these sequences, languages models can be trained to predict masked nucleotides or amino acids and consequently generalise to unseen sequences (Benegas *et al.* 2023; Dee 2022; Lin *et al.* 2023b). Another type of model showing promise on sequential and structural data are generative models. Generative models are self-supervised machine learning models that are trained to model the statistical distribution of input data, typically by reconstructing the original distribution after random noise has been added as input during the training process (Goodfellow *et al.* 2014). A couple of ways in which these models can be applied is to model DNA regulatory sequences (Zrimec *et al.* 2022), and they can be utilised to generate novel protein structures that meet some specified criteria. (Ingraham *et al.* 2023; Watson *et al.* 2023). Attention-based neural networks have shown to be well versed in analysing sequences to correct consensus sequence errors (Baid *et al.* 2023), comprehend protein structures (Baek *et al.* 2021; Lin *et al.* 2023b), and discover potential drug targets (Chen *et al.* 2023). The attention mechanism allows the model to learn relations between different parts of the input sequence, even if these parts are located far away from each other in their representation space (Vaswani *et al.* 2017). The most notable example of an attention-based neural network working with sequence-based data is AlphaFold. AlphaFold predicts protein structure in 3D from an amino acid sequence input (Jumper *et al.* 2021).

Network data (e.g., gene and protein interaction networks) can provide a comprehensive view of molecular relationships, by representing them efficiently as graphs with nodes and edges. Furthermore, representing data as a graph allows us to

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

build Graph Neural Networks (GNNs) (Veličković 2023). GNNs are optimised to learn and propagate information across nodes, allowing for efficient learning from these data structures. In the context of drug target discovery, there are various successful examples of graphs being used, such as in network expansion for pleiotropy mapping (Barrio-Hernandez *et al.* 2023), CausalBench (Chevalley *et al.* 2022), and many others (Muzio *et al.* 2021). A recent trend in drug target discovery has been the usage of Knowledge Graphs (KGs). These typically are heterogeneous graphs that store different data about compounds or genes in nodes, and relationships between nodes in the edges (Chandak *et al.* 2023).

Medical imaging, including x-rays, CT scans, MRI, and histopathology slides, function as important assets for disease diagnosis and tracking treatment responses. Generative models, Convolutional Neural Networks (CNNs), Visual Transformers (ViTs) and deep learning architectures are frequently used for the analysis of visual data (Dosovitskiy *et al.* 2021; Liu *et al.* 2017; Tu *et al.* 2023). When it comes to molecular imaging, images are captured in various resolutions all the way down from the tissue to the cellular level. These images offer profound insights into the molecular intricacies of diseases and drug interactions. Finally, drug screening assays generate a treasure trove of image data, showcasing cells or organisms under perturbation of various compounds in pursuit of potential drugs. AI models help with their ability to comprehensively analyse the resulting images. Next to interpreting the images, using image data also often involves image correction and automatic feature extraction, both tasks in which AI methods excel (Dee *et al.* 2023; Krentzel *et al.* 2023).

While it is true that certain data modalities conventionally have been associated with certain types of AI architectures, a lot of the state-of-the-art models do not exclusively use a single data modality or a single architecture. Often, data and model types are combined. This combination can occur in various ways and often different model types are involved with the processing of various types of data before it gets combined, which often happens in so-called embeddings (Alwazzan *et al.* 2023; Khader *et al.* 2023; Ngiam *et al.* 2011; Venugopalan *et al.* 2021). Embeddings are representations of the raw input data in a latent space that can be used for downstream computations. Furthermore, most modern-day AI architectures consist of various blocks, which are organisational units in a neural network that are composed of different layers, or even

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

whole models that feed into each other and interact with each other. Models like this are often referred to as multimodal machine learning models.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Data modality	Biological representation	Main AI architectures	Example data sources[1]
Tabular	Multiomics, Electronic Health Records	Traditional machine learning[2], Multilayer perceptron	UK Biobank, Genes & Health, OpenTargets, TCGA, GEO
Text	Gene ontology, Scientific literature, Clinical trials	Large language model	GO, PubMed, ClinicalTrials.gov
Sequence/ Structure	DNA, RNA, Protein, Small molecules	Attention-based neural network, Generative model, Language model	Ensembl, UniProt, UCSC Genome Browser, ChEMBL, GenBank, PDB, GENCODE
Graph	PPI, Gene interaction network, Protein structures, Small molecule structures, Pathway annotations	Graph neural network	STRING, STITCH, BioGRID, PDB, TRRUST, RegNetwork, IntAct, PubChem, ChEMBL, Reactome, KEGG
Image	Histopathology, Radiology, Spatial transcriptomics	Convolutional neural network, Visual transformers	TCIA, GDC, MICA-MIC

Table 1: Categorisation of various data modalities commonly used in the field of biomedical research and drug target discovery, along with biology the data represent, the primary AI architecture employed on them, and key data sources. Note that the AI architectures are not exclusive to these data modalities and in practise. Moreover, often multiple are combined or sometimes even integrated into each other in an end-to-end fashion.

[1] Citations to databases can be found in Supplementary Materials S2.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

[2] In this case, we mean traditional machine learning to encompass linear and logistic regression, support vector machines and tree boosting models.

Exploring AI-Based Strategies for Drug Target Identification

The first example we will explore is DrugnomeAI, an ensemble architecture for the prediction of drug targets (Polikar 2006; Raies *et al.* 2022; Vitsios and Petrovski 2020). DrugnomeAI excels in predicting the druggability of candidate drug targets by leveraging 324 gene-level features for every protein-coding gene within the human exome. Raies et al. conducted a feature importance study with Boruta, which is a feature selection technique that helps identify the most relevant variables in a dataset by comparing their importance to that of randomised, noise-added variables (Kursa *et al.* 2010). This analysis showed that the most informative features for druggability prediction were protein-protein interaction features. This is in line with existing research showing that partners of druggable genes are also likely to be druggable (Finan *et al.* 2017). Raies and colleagues frame their model's objective as a positive-unlabelled learning (PUL) problem. Here, the positive dataset comprises targets for which they have identified evidence of druggability, while the unlabelled set encompasses the remaining targets. The ultimate task is to rank these remaining targets based on their predicted druggability. Within their PUL framework, Raies et al. use eight separate classifiers that are stochastically trained through a 10-fold cross-validation process. Subsequently, the predictions from these classifiers are combined to produce the final ranking of the unlabelled drug targets. Notably, Raies et al. observed that the top-ranked genes in their prioritisation exhibit significant enrichment in the clinical literature, arguing that their model has effectively recognised druggability patterns within the feature set.

It is also possible to combine multiple data modalities in a more direct way than ensemble modelling, namely via multitask learning (Caruana 1998). A multitask learning problem in drug target discovery is typically framed as one where you are trying to predict target qualities as well as properties of the target-binding drug (Lin *et al.* 2022; Sadawi *et al.* 2019). Multitask learning allows the model to co-learn a set of tasks together to optimise overall performance. This approach leverages shared information between tasks, combatting overfitting and improving generalisation. Multitask neural networks can integrate data from various sources, making them valuable for a wide range of tasks such as predicting drug targets, but also drug toxicity and sensitivity (Ammad-Ud-Din *et al.* 2017; Costello *et al.* 2014). Furthermore, they offer a means to bridge the gap between biology and chemistry in

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

drug discovery by incorporating structural data like SMILES representations, next to information characterising the biological target, enabling simultaneous prediction of side effects, ligand docking, likely targets, and key compound properties (Mikolov *et al.* 2013b, 2013a).

In some areas of study where data is sparsely available, such as for rare diseases or diseases in clinically unavailable tissues, AI methods can meaningfully identify candidate drug targets through transfer learning. Transfer learning is a concept in AI where we train on abundant data that is tangentially related to some problem with limited data, and consequently fine-tune the resulting model towards the limited data case (Pan and Yang 2010). One example of a model utilising transfer learning is Geneformer (Theodoris *et al.* 2023). Geneformer uses self-attention to pick out important genes using transcriptomic data, which can vary across different cell types, developmental stages, or disease conditions. Geneformer was trained with a dataset called Genecorpus-30M, which was assembled from 29.9 million human single-cell transcriptomes. The transcriptome data is processed through six transformer encoder units, involving self-attention and feed-forward layers. Pre-training is done using a masked learning objective, where 15% of genes in each transcriptome are masked, and the model learns to predict the masked genes based on the context of the unmasked genes. Due to the size and broad scope of Geneformer's pre-training, together with the potential to fine-tune the model, we refer to this model as a foundation model (Bommasani *et al.* 2022). Using Geneformer, cardiomyocytes from three types of limitedly available heart tissue were studied: healthy (n=9), hypertrophic cardiomyopathy (n=11), or dilated cardiomyopathy (n=9). Theodoris et al. performed *in silico* treatment analysis by either inhibiting or activating pathways and seeing if this would move the healthy cell states towards either hypertrophic or dilated cardiomyopathic states. If so, the pathway was inspected for potential therapeutic targets based on druggability and disease relevance. A target that was highlighted through this analysis was *ADCY5*, which is a known druggable target (Wagner *et al.* 2016a) as well as involved in longevity and protection of cardiomyocytes in mouse models (Ho *et al.* 2010). Another target that *in silico* treatment analysis pointed to in this context was *SRPK3*, which is a downstream effector of *MEF2* (Nakagawa *et al.* 2005). *MEF2* is known to play a role in myocardial cell hypertrophy (Akazawa and Komuro 2003). While single-cell foundation models have demonstrated impressive results in certain situations and seem conceptually attractive for downstream applications, it's important to exercise caution. These pre-

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

trained models may not perform well in all contexts, particularly for zero-shot prediction in other biological contexts (Kedzierska *et al.* 2023). Therefore, employing biological foundation models for zero-shot prediction in contexts divergent from their original training objective should be approached carefully.

GNNs are also being employed in drug target discovery. One such approach is EMOGI (Schulte-Sasse *et al.* 2021), a graph convolutional network (GCN) that predicts cancer drug targets. EMOGI stands out by integrating a wide range of interaction and multiomics data to predict cancer genes. This way of combining different data sources addresses the evolving understanding of cancer as a complex interplay of genetic and non-genetic factors (Bell and Gilan 2020; Hanahan and Weinberg 2011). Unlike previous approaches that primarily rely on somatic mutations and occasionally integrate PPI networks (Cowen *et al.* 2017; Leiserson *et al.* 2015; Reyna *et al.* 2018), EMOGI employs GCNs to predict cancer genes by amalgamating multiple data modalities, including mutations, copy number variations, DNA methylation, gene expression, and PPI networks. The graph is constructed to have its topology represent a PPI network. This means that the nodes represent genes, and the edges represent whether two genes interact. R. Schulte-Sassen et al. also did interpretability analysis of their GCN model. They use the Layer-wise Relevance Propagation (LRP) propagation rule (Bach *et al.* 2015), which allows for dissecting what is happening in the GCNs and provides us with insights into why specific genes are classified as cancer-related. Through biclustering and LRP analysis, distinct modules of newly predicted cancer genes (NPCGs) are revealed—some predominantly influenced by network interactions, others primarily driven by omics features. These NPCGs, while not always necessarily displaying recurrent alterations themselves, interact with known cancer genes, positioning them as significant players in tumorigenesis. Notably, these predictions align with essential genes identified through loss-of-function screens, reinforcing the credibility of EMOGI's insights.

Beyond academic research and applications, as of Q3 2023, there are a plethora of AI-derived therapeutics in clinical trial pipelines. Most of these come forth out of industrial research laboratories. A lot of the information that is publicly available on how AI is influencing drug target discovery comes from what we here refer to as AI-first drug discovery companies. These are companies that highlight explicitly the fact that they are using AI in their drug target discovery and drug design efforts. While we

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

can only associate drugs being AI-derived from such companies, we should note that big pharmaceutical companies are also heavily investing into introducing AI into their pipelines. However, it is much harder to attribute the involvement of AI in the development of new pharmaceuticals in this case. So, while looking at the status of AI-first companies might be a good probe into the penetrance of AI into the pharmaceutical industry, it does not provide us with a comprehensive view of the role AI is currently playing in industry.

In Figure 2, we have visualised the status of targets and associated compounds currently in clinical and preclinical trials. The data was put together by searching and collecting a list of publicly and privately held companies that explicitly mention the usage of AI on their website. We have added a table containing the data we collected in Supplementary Table S1. Note that this is not an exhaustive list and we only included target-compound pairs for which we could find sufficient data in the pipelines reported by the companies. For discontinued compounds, press-releases and historical website snapshots have been consulted to confirm the development status of compounds. The discontinued compounds collected in our data is an underestimation of the true number of discontinued compounds. Often, data and status on discontinued compounds is not easily accessible in public records. Hence, the only discontinued compounds added in this list, are ones that (i) have had accessible press coverage, (ii) have been withdrawn from a clinical trial investigation as indicated by ClinicalTrials.gov, or (iii) have been mentioned in an accessible snapshot of a company's pipeline webpage, consulted via wayback.archive.org, and removed without any mention of success. We only consider compounds in which the company was leading the effort for approval. We use FDA approval status to determine whether a compound has been officially approved. We excluded AI-first companies that have not yet had at least one compound enter clinical trials.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

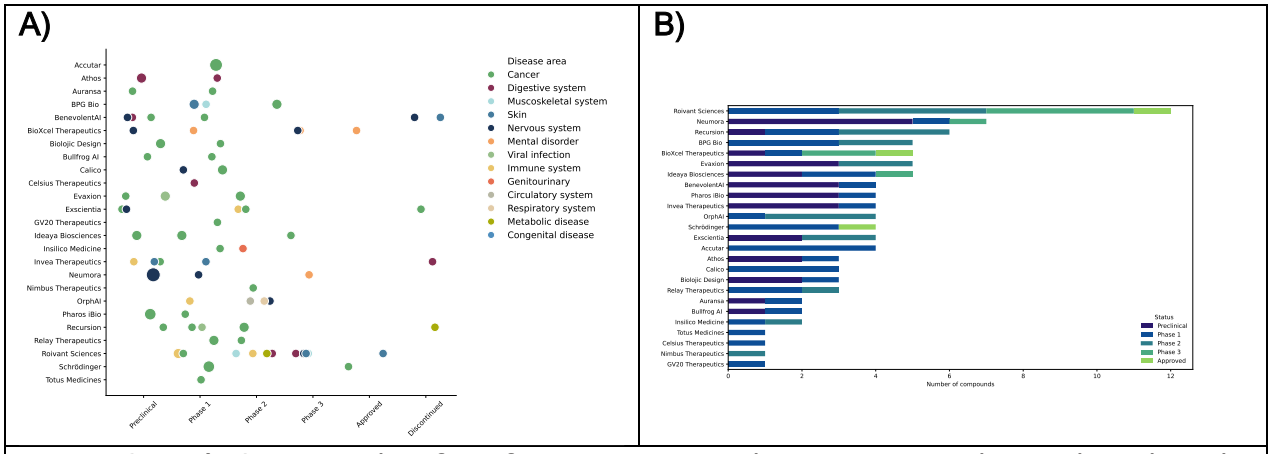

Figure 2. A) Compounds of AI-first companies that are currently in clinical trials, approved or discontinued, stratified by ICD10 disease categories. Scatter size indicates the number of compounds in that clinical trial phase for that company and disease area. Note that dots have been jittered for visual purposes. This does not reflect progress of the compound in the respective phase. B) Number of compounds each company has in clinical trials, where the bar colours refer to the phase or the status of the clinical trial.

Discussion and Future Prospects

AI is penetrating all levels of drug discovery, including target discovery and validation. AI methods rely on the existence of large, high quality data sets. Currently, these data exist but are certainly incomplete and potentially confounding in nature. We must take note of the limitations of existing data and look at ways to improve data in a targeted manner. Most publicly available big data sets often rely on aggregated information descendent from skewed representations of the population. Different populations display widely varying genomic characteristics and responses to drugs, and consequently, less represented populations suffer from diminished treatment outcomes (Gross *et al.* 2022; Popejoy and Fullerton 2016; Ramamoorthy *et al.* 2015). Therefore, the databases used to identify drug targets often lack sufficient representation of population diversity, resulting in disparate health outcomes for diseases that are effectively treated in well-represented groups but remain challenging to address in the underrepresented populations. (Hindorff *et al.* 2018; Landry *et al.* 2018).

At the molecular level, we encounter a different set of biases in the data we use to train our models. For example, some protein classes are significantly overrepresented

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

compared to others based on FDA approval data, which may be attributed to shared structural or functional similarities for proteins within a given class. If we train a new generation of models with these targets as labels, we are likely to perpetuate these biases in newly prioritised drug targets. Furthermore, we should also acknowledge that because of data availability limitations, bias and historical momentum around known drug targets and classes of targets, there is a significant portion of the genome of which we know too little to assess their validity as drug targets (Finan *et al.* 2017; Oprea *et al.* 2018; Wood *et al.* 2019). Assuming there are also potential drug targets hidden within what has been colloquially termed the "unknome" (Rocha *et al.* 2023), this would increase the search space of potential drug targets further beyond what the current paradigm of what drug target druggability models consider. Another challenge is that the concept of a druggable target is not static. This is particularly pronounced for cancer, where target-associated pathways are prone to quickly becoming resistant to treatment through various mechanisms (Shabani and Hojjat-Farsangi 2016). This means that the “one disease, one target” paradigm might not be the best approach to curing diseases, even in cases where a single target is indeed initially therapeutically receptive to treat the disease.

While AI-powered drug target discovery has its fair share of obstacles to overcome, it is still a field that is in its infancy. Moreover, next to these obstacles lie many opportunities for promising discoveries. This is not only limited to drug target discovery, but drug discovery in its broadest sense. For the successful application of AI, specifically deep learning-based architectures, the three guiding principles must be satisfied: scale, complexity, and non-linearity. We argue that drug target discovery satisfies all three of these principles. Given this reality, AI-based methods stand to improve the speed with which we can discover and validate novel drug targets. Recent breakthroughs in AI have led to improvements by providing an increased ability to incorporate sequence and structure-based target evidence. As models like AlphaFold are improved and extended to also reflect the dynamic nature of proteins, and we incorporate small molecules and macromolecular structures into these models, our ability to do *in silico* drug discovery will dramatically improve. In addition to predicting protein structures, AI methods stand to significantly improve a multitude of other biological challenges. These include, but are not limited to, predicting gene perturbations, assessing the effects of genetic variants, *de novo* generation of proteins, and molecular docking simulations. In the long run, transitioning a significant

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

524 portion of the drug discovery pipeline to an *in silico* environment holds substantial
525 advantages for all parties involved with drug discovery. For patients, this shift would
526 enhance the efficiency of developing new and safe medications, resulting in faster
527 delivery of improved therapeutics. For pharmaceutical companies, this transition
528 would lead to significant cost and time savings, which are estimated between 25%
529 and 50% up to the preclinical stage (Loynachan *et al.* 2023). For us to get to this
530 point, experimental validations of *in silico* methods remain essential both to validate
531 computational predictions and to provide labels for the models to train with.
532
533 AI-driven drug target discovery presents a promising avenue for identifying novel,
534 safe and efficacious targets. By leveraging the abundance of multiomics data, and the
535 power of modern AI architectures, applicable to a variety of data modalities – ranging
536 from images to sequences and protein structures – we find ourselves at the precipice
537 of having data and method converge at meaningful impact on drug target discovery,
538 and drug discovery at large.
539

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

References

Abdel-Magid AF (2015) Allosteric Modulators: An Emerging Concept in Drug Discovery. *ACS Medicinal Chemistry Letters* **6**(2), 104–107. https://doi.org/10.1021/ml5005365.

Akazawa H and Komuro I (2003) Roles of Cardiac Transcription Factors in Cardiac Hypertrophy. *Circulation Research* **92**(10), 1079–1088. https://doi.org/10.1161/01.RES.0000072977.86706.23.

Alwazzan O, Khan A, Patras I and Slabaugh G (2023) MOAB: Multi-Modal Outer Arithmetic Block for Fusion of Histopathological Images and Genetic Data for Brain Tumor Grading. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. 1–5. https://doi.org/10.1109/ISBI53787.2023.10230698.

Ammad-Ud-Din M, Khan SA, Wennerberg K and Aittokallio T (2017) Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics (Oxford, England)* **33**(14), i359–i368. https://doi.org/10.1093/bioinformatics/btx266.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25–29. https://doi.org/10.1038/75556.

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P and Kelley DR (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**(10), 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R and Samek W (2015) On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **10**(7), e0130140. https://doi.org/10.1371/journal.pone.0130140.

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ and Baker D (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**(6557), 871–876. https://doi.org/10.1126/science.abj8754.

Baid G, Cook DE, Shafin K, Yun T, Llinares-López F, Berthet Q, Belyaeva A, Töpfer A, Wenger AM, Rowell WJ, Yang H, Kolesnikov A, Ammar W, Vert J-P, Vaswani A, McLean CY, Nattestad M, Chang P-C and Carroll A (2023) DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nature Biotechnology* **41**(2), 232–238. https://doi.org/10.1038/s41587-022-01435-7.

Bakheet TM and Doig AJ (2009) Properties and identification of human protein drug targets. *Bioinformatics* **25**(4), 451–457. https://doi.org/10.1093/bioinformatics/btp002.

Barrio-Hernandez I, Schwartzentruber J, Shrivastava A, del-Toro N, Gonzalez A, Zhang Q, Mountjoy E, Suveges D, Ochoa D, Ghoussaini M, Bradley G, Hermjakob H, Orchard S, Dunham I, Anderson CA, Porras P and Beltrao P (2023) Network expansion of genetic associations defines a pleiotropy map of

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

human cell biology. *Nature Genetics* 1–10. https://doi.org/10.1038/s41588-023-01327-9.

Bell CC and Gilan O (2020) Principles and mechanisms of non-genetic resistance in cancer. *British Journal of Cancer* **122**(4), 465–472. https://doi.org/10.1038/s41416-019-0648-6.

Benegas G, Batra SS and Song YS (2023) DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences* **120**(44), e2311219120. https://doi.org/10.1073/pnas.2311219120.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research* **28**(1), 235–242. https://doi.org/10.1093/nar/28.1.235.

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K and Liang P (2022, July 12) On the Opportunities and Risks of Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2108.07258.

Brandes N, Goldman G, Wang CH, Ye CJ and Ntranos V (2022, August 26) Genome-wide prediction of disease variants with a deep protein language model. bioRxiv, 2022.08.25.505311. https://doi.org/10.1101/2022.08.25.505311.

Bunne C, Stark SG, Gut G, del Castillo JS, Levesque M, Lehmann K-V, Pelkmans L, Krause A and Rätsch G (2023) Learning single-cell perturbation responses using neural optimal transport. *Nature Methods* 1–10. https://doi.org/10.1038/s41592-023-01969-x.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P and Marchini J (2018) The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z.

Caruana R (1998) Multitask Learning. In Thrun S and Pratt L (eds), *Learning to Learn*. Boston, MA: Springer US, 95–133. https://doi.org/10.1007/978-1-4615-5529-2_5.

Chandak P, Huang K and Zitnik M (2023) Building a knowledge graph to enable precision medicine. *Scientific Data* **10**(1), 67. https://doi.org/10.1038/s41597-023-01960-3.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Chen J, Gu Z, Xu Y, Deng M, Lai L and Pei J (2023) QuoteTarget: A sequence-based transformer protein language model to identify potentially druggable protein targets. *Protein Science* **32**(2), e4555. https://doi.org/10.1002/pro.4555.

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*. New York, NY, USA: Association for Computing Machinery, 785–794. https://doi.org/10.1145/2939672.2939785.

Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J and Zhang Y (2016) Drug-target interaction prediction: databases, web servers and computational models. *Briefings in Bioinformatics* **17**(4), 696–712. https://doi.org/10.1093/bib/bbv066.

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, Schneider RG, Senior AW, Jumper J, Hassabis D, Kohli P and Avsec Ž (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**(6664), eadg7492. https://doi.org/10.1126/science.adg7492.

Chevalley M, Roohani Y, Mehrjou A, Leskovec J and Schwab P (2022, October 31) CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data. arXiv. http://arxiv.org/abs/2210.17283 (accessed 8 March 2023)

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L and Prior F (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging* **26**(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Clough E and Barrett T (2016) The Gene Expression Omnibus database. *Methods in Molecular Biology (Clifton, N.J.)* **1418**, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.

Corso G, Stärk H, Jing B, Barzilay R and Jaakkola T (2023, February 11) DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. arXiv. http://arxiv.org/abs/2210.01776 (accessed 30 August 2023)

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, Mpindi J-P, Kallioniemi O, Honkela A, Aittokallio T, Wennerberg K, Collins JJ, Gallahan D, Singer D, Saez-Rodriguez J, Kaski S, Gray JW and Stolovitzky G (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* **32**(12), 1202–1212. https://doi.org/10.1038/nbt.2877.

Cowen L, Ideker T, Raphael BJ and Sharan R (2017) Network propagation: a universal amplifier of genetic associations. *Nature Reviews. Genetics* **18**(9), 551–562. https://doi.org/10.1038/nrg.2017.38.

Dee W (2022) LMPred: predicting antimicrobial peptides using pre-trained language models and deep learning. *Bioinformatics Advances* **2**(1). https://doi.org/10.1093/bioadv/vbac021.

Dee W, Sequeira I, Lobley A and Slabaugh G (2023, December 13) Cell-Vision Fusion: A Swin Transformer-based Approach to Predicting Kinase Inhibitor Mechanism of Action from Cell Painting Data. bioRxiv, 2023.12.13.571534. https://doi.org/10.1101/2023.12.13.571534.

Devlin J, Chang M-W, Lee K and Toutanova K (2019, May 24) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

688 **DiMasi JA, Grabowski HG and Hansen RW** (2016) Innovation in the pharmaceutical
689 industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20-
690 33. https://doi.org/10.1016/j.jhealeco.2016.01.012.
691 **Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T,**
692 **Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N**
693 (2021, June 3) An Image is Worth 16x16 Words: Transformers for Image
694 Recognition at Scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929.
695 **Drews J** (2000) Drug Discovery: A Historical Perspective. **287**(5460), 1960-1964.
696 https://doi.org/10.1126/science.287.5460.1960.
697 **Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, Galver L, Kelley**
698 **R, Karlsson A, Santos R, Overington JP, Hingorani AD and Casas JP** (2017)
699 The druggable genome and support for target identification and validation in
700 drug development. *Science Translational Medicine* **9**(383), eaag1166.
701 https://doi.org/10.1126/scitranslmed.aag1166.
702 **Finer S, Martin HC, Khan A, Hunt KA, MacLaughlin B, Ahmed Z, Ashcroft R, Durham**
703 **C, MacArthur DG, McCarthy MI, Robson J, Trivedi B, Griffiths C, Wright J,**
704 **Trembath RC and van Heel DA** (2020) Cohort Profile: East London Genes &
705 Health (ELGH), a community-based population genomics and health study in
706 British Bangladeshi and British Pakistani people. *International Journal of*
707 *Epidemiology* **49**(1), 20-21i. https://doi.org/10.1093/ije/dyz174.
708 **Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C,**
709 **Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala**
710 **S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C,**
711 **Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG,**
712 **Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker**

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

713 **A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner M-M,**
714 **Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A,**
715 **Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis**
716 **M, Paten B, Tress ML and Flicek P** (2021) GENCODE 2021. *Nucleic Acids*
717 *Research* **49**(D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087.
718 **Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, Gal Y and Marks DS** (2021)
719 Disease variant prediction with deep generative models of evolutionary data.
720 *Nature* **599**(7883), 91–95. https://doi.org/10.1038/s41586-021-04043-8.
721 **Gentile F, Agrawal V, Hsing M, Ton A-T, Ban F, Norinder U, Gleave ME and**
722 **Cherkasov A** (2020) Deep Docking: A Deep Learning Platform for
723 Augmentation of Structure Based Drug Discovery. *ACS Central Science* **6**(6),
724 939–949. https://doi.org/10.1021/acscentsci.0c00229.
725 **Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, Griss J,**
726 **Sevilla C, Matthews L, Gong C, Deng C, Varusai T, Ragueneau E, Haider Y,**
727 **May B, Shamovsky V, Weiser J, Brunson T, Sanati N, Beckman L, Shao X,**
728 **Fabregat A, Sidiropoulos K, Murillo J, Viteri G, Cook J, Shorser S, Bader G,**
729 **Demir E, Sander C, Haw R, Wu G, Stein L, Hermjakob H and D'Eustachio P**
730 (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Research*
731 **50**(D1), D687–D692. https://doi.org/10.1093/nar/gkab1028.
732 **Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville**
733 **A and Bengio Y** (2014, June 10) Generative Adversarial Networks. arXiv.
734 https://doi.org/10.48550/arXiv.1406.2661.
735 **Gross AS, Harry AC, Clifton CS and Della Pasqua O** (2022) Clinical trial diversity: An
736 opportunity for improved insight into the determinants of variability in drug

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

response. *British Journal of Clinical Pharmacology* **88**(6), 2700-2717. https://doi.org/10.1111/bcp.15242.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA and Staudt LM (2016) Toward a Shared Vision for Cancer Genomic Data. *The New England Journal of Medicine* **375**(12), 1109-1112. https://doi.org/10.1056/NEJMp1607591.

Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, Kang B, Jeong D, Kim Y, Jeon H-N, Jung H, Nam S, Chung M, Kim J-H and Lee I (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* **46**(D1), D380-D386. https://doi.org/10.1093/nar/gkx1013.

Hanahan D and Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. *Cell* **144**(5), 646-674. https://doi.org/10.1016/j.cell.2011.02.013.

Hann MM and Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology* **8**(3), 255-263. https://doi.org/10.1016/j.cbpa.2004.04.003.

Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA and Green ED (2018) Prioritizing diversity in human genomics research. *Nature Reviews Genetics* **19**(3), 175-185. https://doi.org/10.1038/nrg.2017.89.

Ho D, Yan L, Iwatsubo K, Vatner DE and Vatner SF (2010) Modulation of β-adrenergic receptor signaling in heart failure and longevity: targeting adenylyl cyclase type 5. *Heart Failure Reviews* **15**(5), 495-512. https://doi.org/10.1007/s10741-010-9183-5.

Huang J, Zhu H, Haggarty SJ, Spring DR, Hwang H, Jin F, Snyder M and Schreiber SL (2004) Finding new components of the target of rapamycin (TOR) signaling

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

network through chemical genetics and proteome chips. *Proceedings of the National Academy of Sciences* **101**(47), 16594-16599. https://doi.org/10.1073/pnas.0407117101.

Huang K, Fu T, Glass LM, Zitnik M, Xiao C and Sun J (2021) DeepPurpose: a deep learning library for drug-target interaction prediction. *Bioinformatics* **36**(22-23), 5545-5547. https://doi.org/10.1093/bioinformatics/btaa1005.

Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, Tie S, Xue V, Cowles SC, Leung A, Rodrigues JV, Morales-Perez CL, Ayoub AM, Green R, Puentes K, Oplinger F, Panwar NV, Obermeyer F, Root AR, Beam AL, Poelwijk FJ and Grigoryan G (2023) Illuminating protein space with a programmable generative model. *Nature* 1-9. https://doi.org/10.1038/s41586-023-06728-8.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P and Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2.

Kanehisa M, Furumichi M, Sato Y, Kawashima M and Ishiguro-Watanabe M (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **51**(D1), D587-D592. https://doi.org/10.1093/nar/gkac963.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Kedzierska KZ, Crawford L, Amini AP and Lu AX (2023, November 5) Assessing the limits of zero-shot foundation models in single-cell biology. bioRxiv, 2023.10.16.561085. https://doi.org/10.1101/2023.10.16.561085.

Kelley DR, Reshef Y, Bileschi M, Belanger D, McLean CY and Snoek J (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research* gr.227819.117. https://doi.org/10.1101/gr.227819.117.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler and D (2002) The Human Genome Browser at UCSC. *Genome Research* **12**(6), 996–1006. https://doi.org/10.1101/gr.229102.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R and Hermjakob H (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Research* **35**(suppl_1), D561–D565. https://doi.org/10.1093/nar/gkl958.

Khader F, Kather JN, Müller-Franzes G, Wang T, Han T, Tayebi Arasteh S, Hamesch K, Bressem K, Haarburger C, Stegmaier J, Kuhl C, Nebelung S and Truhn D (2023) Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports* **13**(1), 10666. https://doi.org/10.1038/s41598-023-37835-1.

Krentzel D, Shorte SL and Zimmer C (2023) Deep learning in image-based phenotypic drug discovery. *Trends in Cell Biology* **33**(7), 538–554. https://doi.org/10.1016/j.tcb.2022.11.011.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Kuhn M, von Mering C, Campillos M, Jensen LJ and Bork P (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research* **36**(Database issue), D684-D688. https://doi.org/10.1093/nar/gkm795.

Kursa MB, Jankowski A and Rudnicki WR (2010) Boruta – A System for Feature Selection. *Fundamenta Informaticae* **101**(4), 271-285. https://doi.org/10.3233/FI-2010-288.

Landry LG, Ali N, Williams DR, Rehm HL and Bonham VL (2018) Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Affairs* **37**(5), 780-785. https://doi.org/10.1377/hlthaff.2017.1595.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* **521**(7553), 436-444. https://doi.org/10.1038/nature14539.

Leiserson MDM, Vandin F, Wu H-T, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, Lawrence MS, Gonzalez-Perez A, Tamborero D, Cheng Y, Ryslik GA, Lopez-Bigas N, Getz G, Ding L and Raphael BJ (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* **47**(2), 106-114. https://doi.org/10.1038/ng.3168.

Lin S, Shi C and Chen J (2022) GeneralizedDTA: combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery. *BMC Bioinformatics* **23**(1), 367. https://doi.org/10.1186/s12859-022-04905-6.

Lin W, Wells J, Wang Z, Orengo C and Martin ACR (2023a, March 20) VariPred: Enhancing Pathogenicity Prediction of Missense Variants Using Protein

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Language Models. bioRxiv, 2023.03.16.532942. https://doi.org/10.1101/2023.03.16.532942.

Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S and Rives A (2023b) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**(6637), 1123-1130. https://doi.org/10.1126/science.ade2574.

Linder J, Srivastava D, Yuan H, Agarwal V and Kelley DR (2023, September 1) Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. bioRxiv, 2023.08.30.555582. https://doi.org/10.1101/2023.08.30.555582.

Liu Y, Chen X, Cheng J and Peng H (2017) A medical image fusion method based on convolutional neural networks. In *2017 20th International Conference on Information Fusion (Fusion)*. 1-7. https://doi.org/10.23919/ICIF.2017.8009769.

Liu Z-P, Wu C, Miao H and Wu H (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, bav095. https://doi.org/10.1093/database/bav095.

Loynachan C, Unsworth H, Donoghue K and Sonabend R (2023) Unlocking the Potential of AI in Drug Discovery. *Boston Consulting Group and Wellcome*.

Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, Becker A, Bennett R, Berry A, Bhai J, Bhurji SK, Bignell A, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Charkhchi M, Cockburn A, Da Rin Fiorretto L, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Ghattaoraya GS, Martinez JG, Guijarro C, Hardy M, Hollis Z, Hourlier T, Hunt T, Kay M, Kaykala V, Le T, Lemos D, Marques-

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Coelho D, Marugán JC, Merino GA, Mirabueno LP, Mushtaq A, Hossain SN, Ogeh DN, Sakthivel MP, Parker A, Perry M, Piližota I, Prosovetskaia I, Pérez-Silva JG, Salam AIA, Saraiva-Agostinho N, Schuilenburg H, Sheppard D, Sinha S, Sipos B, Stark W, Steed E, Sukumaran R, Sumathipala D, Suner M-M, Surapaneni L, Sutinen K, Szpak M, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Walts B, Wass E, Willhoft N, Allen J, Alvarez-Jarreta J, Chakiachvili M, Flint B, Giorgetti S, Haggerty L, Ilsley GR, Loveland JE, Moore B, Mudge JM, Tate J, Thybert D, Trevanion SJ, Winterbottom A, Frankish A, Hunt SE, Ruffier M, Cunningham F, Dyer S, Finn RD, Howe KL, Harrison PW, Yates AD and Flicek P (2023) Ensembl 2023. *Nucleic Acids Research* **51**(D1), D933–D941. https://doi.org/10.1093/nar/gkac958.

Materi W and Wishart DS (2007) Computational systems biology in drug discovery and development: methods and applications. *Drug Discovery Today* **12**(7), 295–303. https://doi.org/10.1016/j.drudis.2007.02.013.

May LT, Leach K, Sexton PM and Christopoulos A (2007) Allosteric Modulation of G Protein-Coupled Receptors. *Annual Review of Pharmacology and Toxicology* **47**(1), 1–51. https://doi.org/10.1146/annurev.pharmtox.47.120505.105159.

Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A and Leach AR (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**(D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

Microsoft Research AI4Science and Microsoft Azure Quantum (2023, November 13) The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4. arXiv. https://doi.org/10.48550/arXiv.2311.07361.

Mikolov T, Chen K, Corrado G and Dean J (2013a, September 6) Efficient Estimation of Word Representations in Vector Space. arXiv. https://doi.org/10.48550/arXiv.1301.3781.

Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J (2013b) Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (accessed 18 September 2023)

Minikel EV, Painter JL, Dong CC and Nelson MR (2024) Refining the impact of genetic evidence on clinical success. *Nature* 1-6. https://doi.org/10.1038/s41586-024-07316-0.

Moffat JG, Vincent F, Lee JA, Eder J and Prunotto M (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature Reviews Drug Discovery* **16**(8), 531-543. https://doi.org/10.1038/nrd.2017.111.

Muzio G, O'Bray L and Borgwardt K (2021) Biological network analysis with deep learning. *Briefings in Bioinformatics* **22**(2), 1515-1530. https://doi.org/10.1093/bib/bbaa257.

Nakagawa O, Arnold M, Nakagawa M, Hamada H, Shelton JM, Kusano H, Harris TM, Childs G, Campbell KP, Richardson JA, Nishino I and Olson EN (2005) Centronuclear myopathy in mice lacking a novel muscle-specific protein

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

kinase transcriptionally regulated by MEF2. *Genes & Development* **19**(17), 2066–2077. https://doi.org/10.1101/gad.1338705.

Ngiam J, Khosla A, Kim M, Nam J, Lee H and Ng AY (2011) Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Madison, WI, USA: Omnipress, 689–696. (accessed 23 November 2023)

Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, Miranda A, Fumis L, Carvalho-Silva D, Spitzer M, Baker J, Ferrer J, Raies A, Razuvayevskaya O, Faulconbridge A, Petsalaki E, Mutowo P, Machlitt-Northen S, Peat G, McAuley E, Ong CK, Mountjoy E, Ghoussaini M, Pierleoni A, Papa E, Pignatelli M, Koscielny G, Karim M, Schwartzentruber J, Hulcoop DG, Dunham I and McDonagh EM (2021) Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Research* **49**(D1), D1302–D1310. https://doi.org/10.1093/nar/gkaa1027.

Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, Jadhav A, Jensen LJ, Johnson GL, Karlson A, Leach AR, Ma'ayan A, Malovannaya A, Mani S, Mathias SL, McManus MT, Meehan TF, von Mering C, Muthas D, Nguyen D-T, Overington JP, Papadatos G, Qin J, Reich C, Roth BL, Schürer SC, Simeonov A, Sklar LA, Southall N, Tomita S, Tudose I, Ursu O, Vidović D, Waller A, Westergaard D, Yang JJ and Zahoránszky-Köhalmi G (2018) Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery* **17**(5), 317–332. https://doi.org/10.1038/nrd.2018.14.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

928 **Paananen J and Fortino V** (2020) An omics perspective on drug target discovery
929 platforms. *Briefings in Bioinformatics* **21**(6), 1937–1953.
930 https://doi.org/10.1093/bib/bbz122.
931 **Pan SJ and Yang Q** (2010) A Survey on Transfer Learning. *IEEE Transactions on*
932 *Knowledge and Data Engineering* **22**(10), 1345–1359.
933 https://doi.org/10.1109/TKDE.2009.191.
934 **Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR and**
935 **Schacht AL** (2010) How to improve R&D productivity: the pharmaceutical
936 industry's grand challenge. *Nature Reviews. Drug Discovery* **9**(3), 203–214.
937 https://doi.org/10.1038/nrd3078.
938 **Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,**
939 **Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D,**
940 **Brucher M, Perrot M and Duchesnay É** (2011) Scikit-learn: Machine Learning
941 in Python. *The Journal of Machine Learning Research* **12**(null), 2825–2830.
942 **Polikar R** (2006) Ensemble based systems in decision making. *IEEE Circuits and*
943 *Systems Magazine* **6**(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199.
944 **Popejoy AB and Fullerton SM** (2016) Genomics is failing on diversity. *Nature*
945 **538**(7624), 161–164. https://doi.org/10.1038/538161a.
946 **Prasad N, Yang K and Uhler C** (2022, January 5) Optimal Transport using GANs for
947 Lineage Tracing. arXiv. https://doi.org/10.48550/arXiv.2007.12098.
948 **Radford A, Narasimhan K, Salimans T and Sutskever I** (2018) Improving Language
949 Understanding by Generative Pre-Training.
950 **Raies A, Tulodziecka E, Stainer J, Middleton L, Dhindsa RS, Hill P, Engkvist O,**
951 **Harper AR, Petrovski S and Vitsios D** (2022) DrugnomeAI is an ensemble
952 machine-learning framework for predicting druggability of candidate drug

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

targets. *Communications Biology* **5**(1), 1–16. https://doi.org/10.1038/s42003-022-04245-4.

Ramamoorthy A, Pacanowski M, Bull J and Zhang L (2015) Racial/ethnic differences in drug disposition and response: Review of recently approved drugs. *Clinical Pharmacology & Therapeutics* **97**(3), 263–273. https://doi.org/10.1002/cpt.61.

Razuvayevskaya O, Lopez I, Dunham I and Ochoa D (2023, February 8) Why Clinical Trials Stop: The Role of Genetics. medRxiv, 2023.02.07.23285407. https://doi.org/10.1101/2023.02.07.23285407.

Reymond J-L (2015) The Chemical Space Project. *Accounts of Chemical Research* **48**(3), 722–730. https://doi.org/10.1021/ar500432k.

Reyna MA, Leiserson MDM and Raphael BJ (2018) Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**(17), i972–i980. https://doi.org/10.1093/bioinformatics/bty613.

Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, Robles C, Freeman M and Munro S (2023) Functional unknomics: Systematic screening of conserved genes of unknown function. *PLOS Biology* **21**(8), e3002222. https://doi.org/10.1371/journal.pbio.3002222.

Royer J, Rodríguez-Cruces R, Tavakol S, Larivière S, Herholz P, Li Q, Vos de Wael R, Paquola C, Benkarim O, Park B, Lowe AJ, Margulies D, Smallwood J, Bernasconi A, Bernasconi N, Frauscher B and Bernhardt BC (2022) An Open MRI Dataset For Multiscale Neuroscience. *Scientific Data* **9**(1), 569. https://doi.org/10.1038/s41597-022-01682-y.

Ruddigkeit L, van Deursen R, Blum LC and Reymond J-L (2012) Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Journal of Chemical Information and Modeling **52**(11), 2864-2875. https://doi.org/10.1021/ci300415d.

Sadawi N, Olier I, Vanschoren J, van Rijn JN, Besnard J, Bickerton R, Grosan C, Soldatova L and King RD (2019) Multi-task learning with a natural metric for quantitative structure activity relationship learning. *Journal of Cheminformatics* **11**(1), 68. https://doi.org/10.1186/s13321-019-0392-1.

Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI and Overington JP (2017) A comprehensive map of molecular drug targets. *Nature Reviews. Drug Discovery* **16**(1), 19-34. https://doi.org/10.1038/nrd.2016.230.

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST and Karsch-Mizrachi I (2021) GenBank. *Nucleic Acids Research* **49**(D1), D92-D96. https://doi.org/10.1093/nar/gkaa1023.

Scannell JW, Blanckley A, Boldon H and Warrington B (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews. Drug Discovery* **11**(3), 191-200. https://doi.org/10.1038/nrd3681.

Schulte-Sasse R, Budach S, Hnisz D and Marsico A (2021) Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence* **3**(6), 513-526. https://doi.org/10.1038/s42256-021-00325-y.

Shabani M and Hojjat-Farsangi M (2016) Targeting Receptor Tyrosine Kinases Using Monoclonal Antibodies: The Most Specific Tools for Targeted-Based Cancer Therapy. *Current Drug Targets* **17**(14), 1687-1703.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A and Natarajan V (2023a) Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Arcas BA y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral J, Webster D, Corrado GS, Matias Y, Azizi S, Karthikesalingam A and Natarajan V (2023b, May 16) Towards Expert-Level Medical Question Answering with Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2305.09617.

Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A and Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**(suppl_1), D535–D539. https://doi.org/10.1093/nar/gkj109.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T and Collins R (2015) UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779.

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ and von Mering C (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Nucleic Acids Research **51**(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000.

Terstappen GC, Schlüpen C, Raggiaschi R and Gaviraghi G (2007) Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery* **6**(11), 891–903. https://doi.org/10.1038/nrd2410.

The UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS and Ellinor PT (2023) Transfer learning enables predictions in network biology. *Nature* **618**(7965), 616–624. https://doi.org/10.1038/s41586-023-06139-9.

Thomas M, Bender A and de Graaf C (2023) Integrating structure-based approaches in generative molecular design. *Current Opinion in Structural Biology* **79**, 102559. https://doi.org/10.1016/j.sbi.2023.102559.

Tu C, Du D, Zeng T and Zhang Y (2023) Deep Multi-dictionary Learning for Survival Prediction with Multi-zoom Histopathological Whole Slide Images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1–12. https://doi.org/10.1109/TCBB.2023.3321593.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I (2017) Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed 4 September 2023)

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Veličković P (2023) Everything is connected: Graph neural networks. *Current Opinion in Structural Biology* **79**, 102538. https://doi.org/10.1016/j.sbi.2023.102538.

Venugopalan J, Tong L, Hassanzadeh HR and Wang MD (2021) Multimodal deep learning models for early detection of Alzheimer's disease stage. *Scientific Reports* **11**(1), 3254. https://doi.org/10.1038/s41598-020-74399-w.

Vitsios D and Petrovski S (2020) Mantis-ml: Disease-Agnostic Gene Prioritization from High-Throughput Genomic Screens by Stochastic Semi-supervised Learning. *The American Journal of Human Genetics* **106**(5), 659–678. https://doi.org/10.1016/j.ajhg.2020.03.012.

Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR, Wilson RK, Mardis ER, Griffith M and Griffith OL (2016a) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Research* **44**(D1), D1036–D1044. https://doi.org/10.1093/nar/gkv1165.

Wagner JR, Lee CT, Durrant JD, Malmstrom RD, Feher VA and Amaro RE (2016b) Emerging Computational Methods for the Rational Discovery of Allosteric Drugs. *Chemical Reviews* **116**(11), 6370–6390. https://doi.org/10.1021/acs.chemrev.5b00631.

Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M and Baker D (2023) De novo design of protein structure and function with

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

RFdiffusion. *Nature* **620**(7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics* **45**(10), 1113–1120. https://doi.org/10.1038/ng.2764.

Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y and Lu H (2017) Deep-Learning-Based Drug-Target Interaction Prediction. *Journal of Proteome Research* **16**(4), 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618.

Wood V, Lock A, Harris MA, Rutherford K, Bähler J and Oliver SG (2019) Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biology* **9**(2), 180241. https://doi.org/10.1098/rsob.180241.

Zrimec J, Fu X, Muhammad AS, Skrekas C, Jauniskis V, Speicher NK, Börlin CS, Verendel V, Chehreghani MH, Dubhashi D, Siewers V, David F, Nielsen J and Zelezniak A (2022) Controlling gene expression with deep generative design of regulatory DNA. *Nature Communications* **13**(1), 5099. https://doi.org/10.1038/s41467-022-32818-8.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Impact Statement

Artificial intelligence (AI) is transforming drug discovery and development by enabling the rapid analysis of massive amounts of biological data and chemical information. This paper reviews recent advances in using AI methods for the discovery and validation of drug targets. Identifying and validating novel drug targets is fundamental to creating safe and effective new medicines but has remained a major bottleneck in the drug R&D process. By integrating diverse datasets, AI models can accurately predict key properties of drug targets, reveal intricate biological relationships underlying disease, and guide drug discovery strategies.

This paper highlights groundbreaking applications of AI that accelerate target discovery, including models that prioritise candidate genes, predict druggability of proteins, uncover disease mechanisms, and simulate biological experiments. Critically, AI enables leveraging insights across modalities like sequences (e.g. DNA, proteins), structures (e.g. compounds, proteins), multiomics, biomedical literature and more. Integrating multimodal inputs is paramount for comprehensively understanding complex diseases involving genetic and non-genetic factors.

The AI methods outlined will profoundly enhance R&D efficiency. By illuminating novel drug targets, AI-powered target discovery will expand treatment options available for patients suffering from previously untreatable or poorly managed diseases. From rare diseases and refractory cancers to multifactorial neurodegenerative and autoimmune conditions, accelerating target discovery through AI has far-reaching therapeutic implications. Additionally, safer, more selective drugs developed against AI-predicted targets could dramatically improve patient outcomes and quality of life. Overcoming existing challenges in AI-based target discovery will be critical to actualising its immense potential and promises to usher in a new era of data-driven, accelerated drug R&D.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press

Financial Support

C. C. was funded by the National Institute for Health Research (NIHR) as part of the portfolio of translational research of the NIHR Biomedical Research Centre at Barts and The London School of Medicine and Dentistry. A. W. was funded by the UKRI/BBSRC Collaborative Training Partnership in AI for Drug Discovery and Queen Mary University of London.

Conflict of Interest

At the time of writing, W. W. and V. N. were employed by MSD.

https://doi.org/10.1017/pcm.2024.4 Published online by Cambridge University Press