Cell Reports Methods

CellPress
OPEN ACCESS

Perspective

Machine learning approaches to predict drug efficacy and toxicity in oncology

Bara A. Badwan,[1] Gerry Liaropoulos,[1] Efthymios Kyrodimos,[2] Dimitrios Skaltsas,[1] Aristotelis Tsirigos,[3,4,*] and Vassilis G. Gorgoulis[1,5,6,7,8,*]

[1]Intelligencia Inc, New York, NY 10014, USA
[2]First ENT Department, Hippocration Hospital, National Kapodistrian University of Athens, Athens, GR 11527, Greece
[3]Department of Medicine, New York University School of Medicine, New York, NY 10016, USA
[4]Department of Pathology, New York University School of Medicine, New York, NY 10016, USA
[5]Department of Histology and Embryology, Faculty of Medicine, School of Health Sciences, National Kapodistrian University of Athens, Athens 11527, Greece
[6]Ninewells Hospital and Medical School, University of Dundee, Dundee DD1 9SY, UK
[7]Biomedical Research Foundation, Academy of Athens, Athens 11527, Greece
[8]Molecular and Clinical Cancer Sciences, Manchester Cancer Research Centre, Manchester Academic Health Sciences Centre, University of Manchester, Manchester M20 4GJ, UK
*Correspondence: aristotelis.tsirigos@nyulangone.org (A.T.), vgorg@med.uoa.gr (V.G.G.)
https://doi.org/10.1016/j.crmeth.2023.100413

SUMMARY

In recent years, there has been a surge of interest in using machine learning algorithms (MLAs) in oncology, particularly for biomedical applications such as drug discovery, drug repurposing, diagnostics, clinical trial design, and pharmaceutical production. MLAs have the potential to provide valuable insights and predictions in these areas by representing both the disease state and the therapeutic agents used to treat it. To fully utilize the capabilities of MLAs in oncology, it is important to understand the fundamental concepts underlying these algorithms and how they can be applied to assess the efficacy and toxicity of therapeutics. In this perspective, we lay out approaches to represent both the disease state and the therapeutic agents used by MLAs to derive novel insights and make relevant predictions.

INTRODUCTION: THE NEED FOR MACHINE LEARNING AND MATHEMATICAL MODELING IN BIOMEDICINE

Machine learning algorithms (MLAs) are a set of algorithms within the field of artificial intelligence (AI) that can learn relevant relationships within large datasets and develop ideal approaches to their analysis without prior specification.[1–4] MLAs have found many applications in drug development, including FDA approval predictions, clinical trial design, drug repurposing, and even generation of new therapeutic targets.[1–4] The field has experienced a rapid development in the past decade and is now reaching a degree of maturity and sophistication that is continually improving.

In the following sections we discuss the basics of MLAs and lay out a framework for how they can be used for drug development. We focus on the methods that have been developed for creating representations of both the therapeutics of interest as well as the disease to be targeted. Then we present the models that leverage these representations to predict the efficacy and toxicity of new therapeutics.

The field of oncology has been a particular focus for the development of new therapeutics and key advances in machine learning (ML) technology have occurred within the cancer context.[3–5] We delve into the details and highlight the resources available, principally in this field of research.

We outline the general approach underlying MLA models in the therapeutics domain, as presented in Figure 1, focusing mainly on models to predict the efficacy and toxicity of new therapeutics, which in turn inform their likelihood of approval. Particularly, we summarize this layout by showing key features, model types, and the insights that each can provide (depicted in detail in Figure 2). In terms of features, we show in Figure 2A that they can be split into two key domains: therapeutic and disease state representations, respectively. In the top-left panel, we focus on the small-molecule and protein therapeutic types and show their innate structure and the various methodologies that have been developed to represent them. For the disease state representation, we summarize the related -omic profiles and their corresponding analyses in the bottom panel. Next we demonstrate, as depicted in Figure 2B, the types of models with which both feature types can be utilized either separately or together. Specifically, we highlight the key model types in both the supervised and the unsupervised domains. Finally, we highlight (in Figure 2C) the different predictions or insights each of those models can generate. The predictions can be characterized as either drug assessment or drug design. For assessment models, a therapeutic entity is pre-defined and the value to be predicted is its potential efficacy or toxicity. For drug design, the models themselves would generate potential therapeutics for a particular disease state. Generative autoencoders can be trained on existing

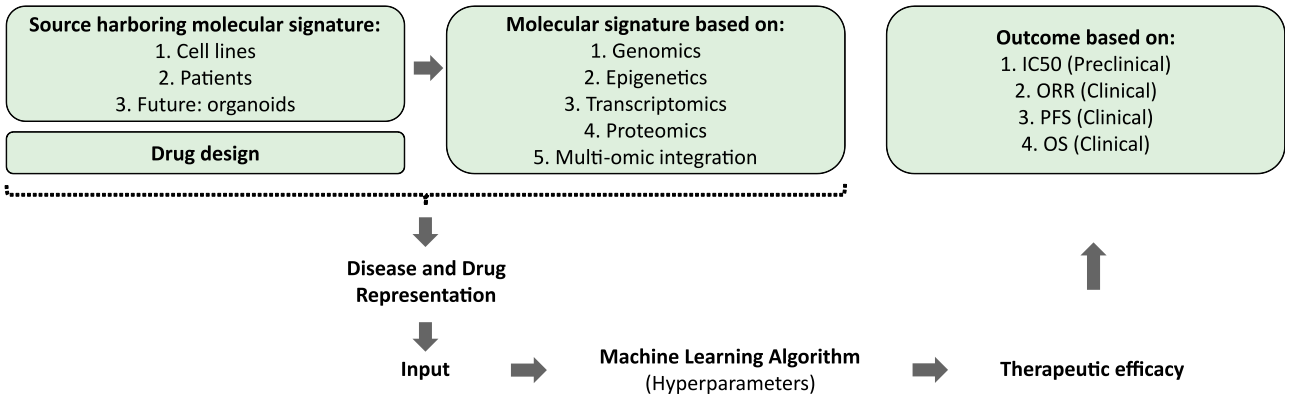

Figure 1. A general outline representing a machine learning algorithm dealing with drug response efficacy in patients with cancer
For details, see sections "the landscape of machine learning," "representations of drug molecules," "representations of disease states," and "therapeutic efficacy," as well as Figure 2.

drugs and their efficacy and toxicity can be used to generate new examples of therapeutics that would be safe and efficacious.[6]

THE LANDSCAPE OF ML

In this section, we give a brief overview of the types of ML and artificial models that have emerged broadly in the past decade. A more detailed exploration of each is provided in our previously published work.[3,7] The first distinction we make is between supervised and unsupervised learning.

Supervised learning

In supervised learning, we generally utilize a large dataset of labeled data to develop a model that is capable of classifying new entries with the correct label. It has broad applications for drug discovery and design as it can be used to assess the efficacy, toxicity, and likelihood of approval of a new therapeutic. Here, we lay out the basics of the approach to inform the discussion in the rest of the review.

Bias-variance tradeoff

In any supervised learning approach there is an underlying tension known as the bias-variance tradeoff, which emerges from two primary concerns that must be accounted for. One relates to the insufficient relevant data to generate valid rules (the bias error) and second that the rules generated are too specific to the particular dataset that is being used to train on (the variance error).

The bias error can be thought of as *underfitting* and refers to an algorithm missing the relevant relationships between the features of interest and what is being predicted. The variance error, also known as *overfitting*, can be considered as the sensitivity of the algorithm to changes in the data, where a model is able to make very accurate predictions with the dataset it was trained on, but fails to generalize and performs poorly on new data. Understanding the trade-off will elucidate the rationale behind the approaches taken when developing these models.

Train-test splitting

The train-validate-test is one of a number of standard approaches that have emerged to deal with the bias-variance tradeoff. The train-validate-test approach requires the splitting of the dataset into two major subsets, the "train set" and the "test set." The former can then be further split into a true train set and a validation set. The model is then trained on the train set, and its performance assessed on the validation one. The hyperparameters (model parameters that are set prior to training as opposed to ones derived via training) of the model can then be adjusted to improve the performance on the validation set. Once the training and tuning is optimized, then the model can be assessed on the test set that had not been used in any way. The approach is meant to avoid the possibility of overfitting the model to the specifics of the data being used and as a result can create a generalizable model that would perform well on previously unseen data.

k-fold cross-validation

A major consideration with the train-test approach is whether or not the split is done in a truly random fashion and if the resulting subsets are appropriately representative. The method of k-fold cross validation builds upon the train-test approach by running the split multiple times. In k-fold cross validation, the train set is split into a k-subsets, and one of the subsets is held out and used for validation; this process is repeated k times with a different subset being held out each time. The performance is then taken to be the average of the k models trained.

There are also methods of introducing a degree of stochasticity into the training process by including slight variations in the datasets used for training or including dropout layers (a ML technique where certain neurons are ignored during training in a stochastic fashion), where certain learned processes are randomly inhibited allowing for the development of more flexible programs that have a better chance at being truly generalizable.[7]

Classification or regression modeling

Supervised learning models can be further categorized based on the type prediction they are making. The two major model types are regression and classifiers.

In a classifier model, the prediction of interest takes one of a few discrete values (e.g., 0 or 1 in a binary manner). A model to assess whether a drug will be approved or rejected would be

CellPress
OPEN ACCESS

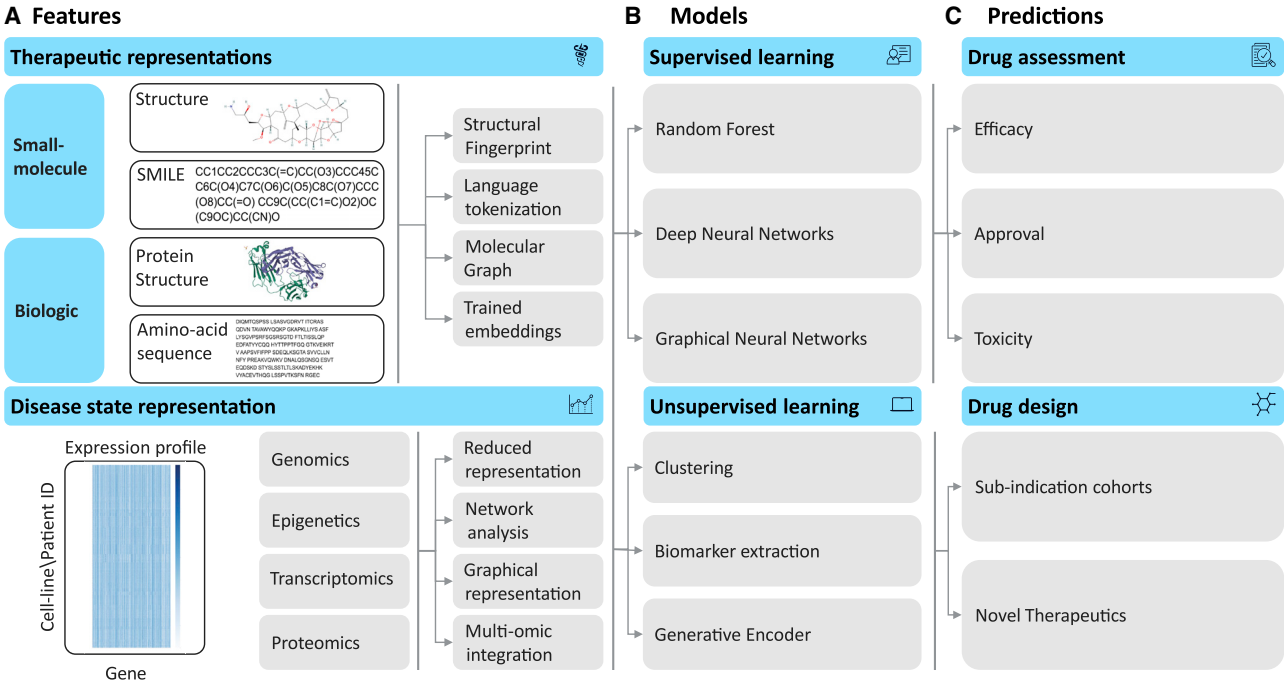

Figure 2. Overview of preclinical efficacy model and patient genomic profiles

(A) Features comprise. (Top) Therapeutic representations: the structure and baseline representation for small-molecule and biologic therapeutics. The molecular structure of eribulin (an anticancer chemical drug, brand name Halaven) and its SMILE structure, and the protein structure of trastuzumab (an anticancer therapy monoclonal antibody, brand name Herceptin) and its amino acid are depicted, respectively. Listing of the possible approaches to generate informative therapeutic representations. (Bottom) Disease state representation: the RNA expression profile of cancer cell lines from CCLE (https://sites.broadinstitute.org/ccle/). The four -omic profiles and their associated processing algorithms are listed.

(B) The models. Listing of the supervised learning model used for drug assessment: random forests, deep neural networks, and graphical neural networks. Listing of the unsupervised learning model used for drug design: clustering, biomarker extraction, generative autoencoders.

(C) The predictions comprise. The drug assessment models that can be used as regressor models to predict the efficacy and toxicity models, or as classifier models to predict approval. The drug design unsupervised models can be used to find new cohorts of patients who could respond to treatments differently, and generate novel drugs to treat specific patient cohorts.

considered as such. Regarding regression models, the prediction of interest can take on any continuous value. If we wanted to predict the efficacy of a cancer drug, by measuring the dosage of the drug required to inhibit 50% of the cancer cells *in vitro* (the IC_{50}), we would use a regression model.

Certain ML models are used exclusively for either classification or regression. For example, linear regression should only be used for regression models, whereas logistic regression, despite the name, should only be used for classification problems. Certain models such as decision trees, random forests, support vector machines, k-nearest neighbors, and neural networks, have classifier and regressor versions and can be used for either problem type.

Unsupervised learning

In unsupervised learning the dataset used as a starting point is unlabeled and the models used are intended to reveal insights into the underlying structure of the data. The primary outcomes are: (1) dimensionality reduction, (2) data visualization, (3) feature extraction, and (4) clustering. These algorithms vary widely in terms of approach and outcomes, and we review its core concepts in the following sections.

Dimensionality reduction

In dimensionality reduction approaches, highly dimensional data (e.g., a transcriptomic profile of 20,000 genes for 10,000 patients as an example) are condensed into their most informative dimensions (e.g., in 2 dimensions for each patient). There are a number of ways to distill the most important dimensions from such a dataset. These techniques have been reviewed previously[3,7–9] and include (1) principal-component analysis (PCA), (2) t-distributed stochastic neighbor embedding (t-SNE), (3) linear discriminant analysis, (4) Uniform Manifold Approximation and Projection. Each of the algorithms has a unique approach, but the underlying concept is the same. Higher-order data are reduced into a smaller set of dimensions, which can then be used for either visualization, feature extraction, or as predictive components in other MLA models.[7]

A clear application of dimensionality reduction in drug discovery is the use of these techniques for high-dimensional patient data such as RNA sequencing (RNA-seq) expression. To demonstrate the process of unsupervised learning we show in Figure 3A how training a t-SNE model on unlabeled cancer patient genomic information results in the natural emergence of clusters that correlate with the type of cancer the patient had.

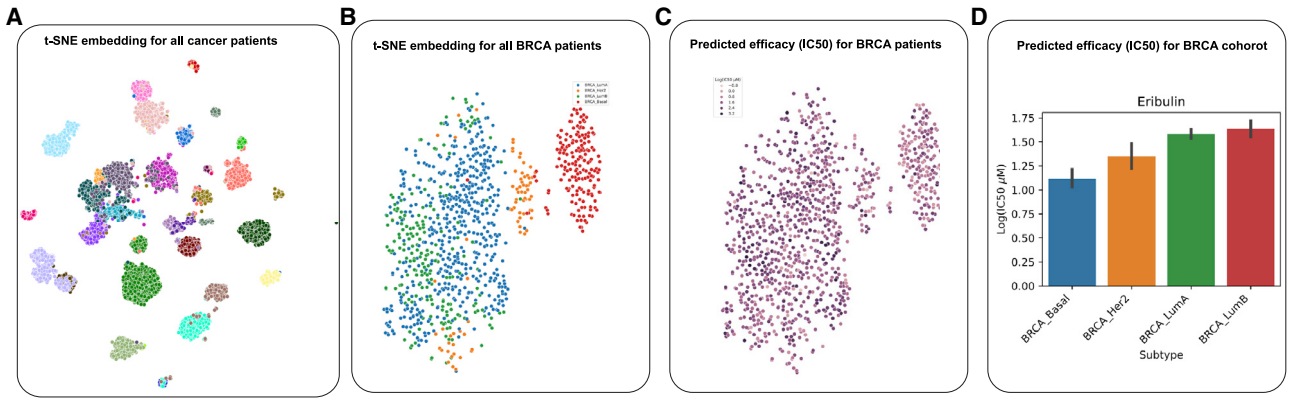

Figure 3. Dimensionality reduction and efficacy prediction for cancer patients

(A) The RNA-seq expression profiles of cancer patients were clustered using t-SNE embedding. Each patient's profile is presented as a single dot in the reduced dimension. Each patient is color labeled according to their TCGA cancer type.

(B) Same as in (A) but focused on breast cancer (BRCA) patients with each patient colored by their BRCA subtype.

(C) The predicted response of each patient to eribulin according to the PaccMann IC_{50} prediction model[6] utilizes the genome expression profile and the chemical structure of the drug to predict efficacy (see section "models of interest to predict efficacy," subsection"clinical efficacy modeling").

(D) The predicated efficacy averaged according to the breast cancer subtypes (see section "models of interest to predict efficacy," subsection"clinical efficacy modeling"). Error bars connote the 95% confidence intervals range.

In Figure 3B, t-SNE is used on the transcriptomic profiles of 1,086 breast cancer (BRCA) patients acquired from TCGA (The Cancer Genome Atlas Program, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga; see Table 1). The transcriptomic profile consists of the gene expression values of 17,715 genes. After the t-SNE analysis, the 17,715 genetic dimensions are reduced to just 2, allowing us to easily visualize the data as the two-dimensional plot shown. Each point here is a tumor sample from a BRCA patient. The distance between the points is an indicator of the degree of similarity between the patient samples. We label each sample with the identified BRCA subtype according to TCGA. We find that the clusters correspond well to the indicated cancer subtype.

Clustering techniques

Clustering algorithms can be used to find large-scale structures within a dataset.[7] These algorithms can split the data points in a dataset into a specified number of clusters and assign each point to one of the clusters. Clustering can reveal higher-order structures within the dataset and help determine similarity between different entries.

Types of such algorithms include[3,7,21–24] (1) k-means clustering, (2) hierarchical clustering, (3) Fuzzy C means clustering, (4) mean shift clustering, (5) density-based spatial clustering of applications with noise, and (6) Gaussian mixed models. Moreover, they can reveal mislabeling within certain datasets, where entries supposedly belonging to one group are revealed to belong to another. Clustering can be especially useful in the context of drug design as it can reveal patient sub-populations that might be more or less sensitive to particular treatment regimes.

Neural network encoders

Autoencoders are a relatively new form of unsupervised learning models that learn to generate data that resemble the input data they are presented with.[25] The data are fed into a neural network, and then regenerated from the reduced embedding that the neural network develops. These models are called generative models (a form of neural networks that learn to create new examples of the data types that are used to train them) as they create new data points in accordance with the specifications of the input data. In drug discovery, these models can be used to generate new possible therapeutics with the requirement of having certain efficacy and toxicity profiles.[6]

REPRESENTATIONS OF DRUG MOLECULES

Constructing an ML algorithm to connect a molecular state, reflecting a disease, with the response to a particular therapeutic intervention and more specifically to the actual drug molecule, faces certain challenges. One of them is to select the best computer-readable form to represent the therapeutic agent under investigation. Here, we discuss the major approaches developed to address this question (see also Figure 2).

Table 1. List of the major bioinformatic databases referenced in the text

Database name	Data types
GTEX[10]	genomic/transcriptomic
UK Biobank[11]	genomics
TCGA[12]	genomics/transcriptomics
ENCODE[13]	epigenetics
STRING[14]	proteomics
ProteomicsDB[15]	proteomics
KEGG[16]	genomic/pathway
REACTOME[17]	pathway
PubChem[18]	small-molecule properties
UniProt[19]	protein properties
ChEMBL[20]	therapeutics database
clincaltrials.gov	clinical outcomes

CellPress
OPEN ACCESS

Small-molecule representation

A small molecule is generally defined as an organic compound with a molecular weight of less than 500 Da.[26] The manageable size of small-molecule drugs allows for a tractable representation of their structures in a computer-readable way. In this section we review the various methods that are available for representing small molecules in a computer-readable manner.

SMILE

One approach of drug structure representation is the simplified molecular input line entry system (SMILE).[27] The chemical annotation system uses a few syntactical rules to allow for the representation of a molecular structure in a computer-readable form. SMILE structures use characters to represent each of the atoms within a molecule and special ones to represent the bonds between them as well as the higher-order structural properties of the molecule such as aromaticity or cyclicality.

Interest in using SMILEs in the context of ML and generative models revealed a major problem. The generated SMILEs might not correspond to valid molecules. Addressing this issue led to the development of the self-referencing embedded strings (SELFIES),[28] which modifies the initial system to ensure that all generated strings refer to valid chemical molecules. Neither SMILES nor SELFIES can be directly used in ML models, as they often require their inputs to be in a vectorized or numerical form, whereas SMILES are character representations. Multiple approaches have emerged to confront this issue.[29]

Fingerprinting

One method of embedding SMILE structure is called fingerprinting, where a chemical structure is converted into a binary vector of pre-determined size that captures the structural information of the original compound. One of the most utilized fingerprinting techniques is Morgan fingerprinting.[30] Binarizing the chemical structure allows for the utilization of model architectures that expect binary vector input. Other fingerprinting techniques have been developed since to expand the capacity of and improve upon Morgan fingerprinting.[31] Vectorizing the molecular structure of a therapeutic through fingerprinting makes it possible to leverage a number of ML architectures that require numerical features.

Natural language processing

With advances in natural language processing (NLP) models, an NLP approach to chemical structure embedding has gained traction in recent years. In this context, the SMILE/SELFIE string is tokenized and a specific language is trained to embed the chemical structure.[6]

Utilizing NLP-inspired models allows the models to capture higher-order relationships across larger distances within the molecule of interest. The NLP approach has been found to outperform the fingerprinting technique in a number of different classification tasks.[32] However, these NLP techniques are still somewhat underutilized, thereby providing a ripe area for researchers in the field to improve the models and predictions.

Molecular graphs

Graphical representations of molecules are another way to capture the full complexity of the molecular therapeutic. In this framework, each atom is encoded as a node in a graph and the connections between them constitute edges. Creating molecular graphs has become a routine operation that can be easily conducted through software modules such as RDKit in Python, where the Le Verrier-Faddeev-Frame approach is applied.[33] The use of graphs to represent molecular structures has become a standard feature of many top-of-the-line drug efficacy models.[34–36] However, they do require additional complexity in terms of the architectures of the models that can utilize them. Therefore, they are better suited for larger therapeutics, such as proteins and peptides, where fewer adequate alternatives exist.

Protein/peptide representation

Representing protein therapeutics in a computer-readable form poses significant challenges that are not present with small molecules. The size and complexity of a protein therapeutic makes the previous approaches untenable.

Protein sequences can be embedded by their physical properties or by their amino acid sequences.[37] Using physical properties poses a challenge as it is difficult to know *a priori* which properties will be most relevant to a learning task. Multiple methods of embedding the amino acid structures have been developed in the past decade and are reviewed below.

NLP for protein encoding

NLP approaches such as word2vec and doc2vec have been used to develop learned embeddings of words or sentences based on their context and surrounding words.[38,39] A number of attempts have been made to apply these approaches to protein sequences by segmenting the protein sequences into fragments of length k (k-mers).[40–43] The protein embedding then learns which segments of a protein sequence are expected to appear next to one another. The approach can then be combined with task-specific learning to create embeddings that learn to extract the relevant aspects of the amino acid sequence.

Task-assisted protein embeddings

This is an approach building upon the NLP and a semi-supervised task ML paradigm described above.[44] Task-assisted protein embeddings (TAPE) utilizes biologically relevant tasks to create an informed protein embedding from an amino acid input. The tasks highlight three major areas of protein biology: (1) structure prediction, (2) detection of remote homology, and (3) protein engineering. Rather than utilizing the word2vec or doc2vec approach, the TAPE approaches utilize other NLP paradigms, namely next-token prediction and masked-token predictions.[45,46] The TAPE embeddings have been adopted widely and have been used in a number of higher-order models such as IBM's PaccMannRL.[6]

Graphical representations

Graphical protein representations have been developed and have been quite successful in predicting protein function and interactions.[47] In these graphs, each node is an amino acid residue and the edges contain information regarding the distances and angles between residues.[48] Such a representation scales more efficiently compared with 3D structural representations used in convolutional neural nets.[49]

REPRESENTATIONS OF DISEASE STATES

The previous sections described the work conducted to develop representations of the therapeutic agent. The ML models of interest also require a representation of the disease state, the

Table 2. Performance of various deep neural networks models for IC_{50} prediction using different therapeutic and disease-state representations

Model Name	Therapeutic representation	Disease state representation	Spearman correlation
DrugCell[5]	Morgan fingerprint	genomic	0.80
Paccmann[6]	SMILE language tokenization	transcriptomic	0.88
DeepDSC[55]	Morgan fingerprint	transcriptomic	0.84
CDRScan[56]	molecular fingerprint	genomic	0.84
tCNNs[57]	one-hot encoding	genomic	0.84
GraphDRP[34]	molecular graph	genomic	0.85
DeepCDR[35]	molecular graph	genomic and transcriptomic	0.82
AGMI[36]	molecular graph	genomic and transcriptomic	0.92

Spearman correlations were reported in the papers referenced or in.[36]

therapeutic intends to target. The classic approach is to think of the disease representation in terms of a genetic or protein target that is associated with disease progression, which the drug would interact with.

The early interest in ML-assisted drug design focused on the intersection of molecular dynamic modeling with ML being utilized to design therapeutic molecules specifically targeting a disease-associated enzyme's active site.[50] The representation of a disease state as a single gene or protein target of interest has been covered extensively elsewhere[51,52] and can be best appreciated in the context of the quantitative structure-activity relationship,[53] which will not be covered here. Instead, we center the higher-order representations profiling the disease state to include the genomic, epigenetic, transcriptomic, and proteomic profiles of the diseased cell, either *in vitro* or from patients suffering from a specific disease (Figure 1). We will look at these approaches specifically in oncology and consider how they might be extended to other indications.

Genomics

The genomic profile of a disease state can be identified through the genetic sequencing of patients or disease state models. The genetic sequence allows for the identification of key mutations that are present and may differentially affect the onset of the disease and the outcome possibilities. Genomic mutations can be a single-nucleotide variant or single-nucleotide polymorphisms, insertions, deletions, inversions, copy number variations, tandem duplications, dispersed duplications, mobile element insertions, or translocations. The genomic mutational profile can then be used as a feature for ML models.[5] The mutational status and copy number variation have been used repeatedly to predict the potential efficacy of new therapeutics and are summarized in Table 2.[34,36,54]

Epigenetics

Epigenetic modifications are critical to gaining a full understanding of the processes underlying a biological state. Comprehensive databases for epigenetic information are currently being developed and are a fast-growing field in bioinformatics.

One highly informative structural feature that can provide epigenetic insights is accessible chromatin. Human assay for transposase-accessible chromatin with high-throughput sequencing (a method to assess genome-wide chromatic accessibility datasets) provides a detailed map of accessible chromatin, has been accumulating rapidly in recent years, and an effort has been undertaken to provide annotated data in a centralized publicly accessible database.[58]

Beyond that, the Roadmap Epigenomics Mapping Consortium project, as part of the Encyclopedia of DNA Elements (https://www.encodeproject.org/), has gathered information on DNA methylation, histone modification, chromatin accessibility, and small RNA transcripts in primary human tissues.[13,59]

The epigenetic tracks provided by the Roadmap Epigenomics Mapping Consortium have been used to train a convolutional neural network to predict mutational rates within genomic regions, and find mutations that have positive associations with sub-cancer types.[60]

Transcriptomics

One of the most ubiquitous -omic profiles used in computational bioinformatics today is the transcriptomic profile, which is captured through RNA-seq expression data. Here, the degree of mRNA expression gives a sense of which genes are activated and which are inhibited in a given cell.

RNA-seq profiling is conducted on a bulk population of cells or in single cells. High-throughput sequential RNA-seq can also allow for spatiotemporal sequencing showing how the mRNA expression profile shifts over time or across spatially separated cells.

Proteomics

Databases of protein structure, properties, interactions, and abundances all inform the proteomic profile of the diseased state. The structural properties and amino acid sequences of proteins found in UNIPROT[19] are used to create reduced embedding of protein targets and biologic therapeutics.

The CHEMBL database (https://www.ebi.ac.uk/chembl/) provides key features and ontologies for antibodies and therapeutically relevant protein targets, which can be used as features in drug prediction models directly, or to create protein networks and similarity metrics for possible drug targets.

The ProteomicsDB (https://www.proteomicsdb.org/) provides mass spectrometry data determining protein abundances in different biological tissue,[15] providing a proteomic profile for the disease state, which can be combined with the other -omic profiles described.

MOLECULAR PATHWAYS AND INTERACTIONS

Dedicated databases capture the interactions between individual genes, transcription factors, mRNA, and proteins as biological pathways. Reactome,[17] KEGG,[16] the Pathway Commons,[61] and Omnipath[62] are major databases that catalog biological pathways. They can be used to construct genomic networks to create disease signatures, and to find with pathways are

particularly affected in the diseased state. The STRING database (https://string-db.org/) provides information on both physical and functional protein-protein interactions (physical contacts of high specificity between two or more proteins),[14] which can be used with network propagation algorithms to find genomic signatures of interest and reduce the dimensional complexity of -omic data generally.[63] These databases can be leveraged and integrated to create a holistic view of the biology underlying the diseased state.

While each of the -omic data types described above can be used independently to predict drug response, models that combine multiple data types have been found to yield more accurate results.[64,65] Various architectures of combining clinical and genomic data for cancer patients have been developed. One approach is to use autoencoders to condense different data types into a reduced embedding and then combine the embeddings themselves.[66] Another is COSMOS (causal oriented search of multi-omic space), an -omic integration method that systematically generates mechanistic hypotheses through causal reasoning.[64,65] COSMOS generates *trans*-omic networks that capture the relationships between entities across -omic levels.

The *trans*-omic networks are used to find signatures, or fingerprints, of disease subtypes. Gene signatures allow researchers to use a smaller subset of genes as key markers, reducing the complexity of the -omic profiles generated.[67]

Knowledge graphs

Another method to combine multiple data types is to use a knowledge graph embedding (KGE), reflecting the disease state.[68] Multiple, specific reviews have covered the subject of KGE recently.[69,70] Knowledge graphs are heterogeneous, which sets them apart from homogeneous graphs by the fact that the edges and nodes can be of differing types.

With this approach, the therapeutic and -omic profiles are embedded as entities features in a graph, and the interactions of the different entities are expressed as relations. The -omic relationships are captured through the following data types: (1) gene ontologies (a formal representation of the body of knowledge within the genomic domain), (2) gene-gene interactions (a set of functional associations between genes), (3) protein-protein interactions, (4) gene pathways (sequential steps that are mediated by gene function that operate together to determine a biological process), and (5) Pearson correlation coefficients (a measurement of the degree of similarity between two entities).

Knowledge graphs can be considered as a series of triplet structures that describes the relationship r between two entities, e_1 and e_2. The entities could refer to genes, therapeutics, or even broader biological concepts. For example, appearing as (gene A, *regulates*, gene B) or (disease A, *downregulates*, gene B), and even (drug A, *treats*, disease A). The relational datasets are noisy and incomplete, where the relationship may appear as (disease A, *downregulates*, ?), or (drug A, *treats*, ?), or (?, *treats*, disease A). The drug discovery process can then be reformulated as finding missing links between the various embedded entities. Prediction models can be trained to find these missing links, which in turn may lead to finding new disease biomarkers, drug repurposing, and drug discovery, respectively. The network representations generated could then be used for discovery of disease gene signatures.[67]

Possible KGE model architectures include: ComplEx,[71] DistMult,[72] RotatE,[73] TransE,[74] and TransH.[75] Hetionet[76] and BioKG[77] are KGE model architectures developed specifically for drug discovery.

THERAPEUTIC EFFICACY

To monitor the therapeutic efficacy, we need measures of effectiveness for a therapeutic. Various values are used and discussed below.

Preclinical IC_{50}

The cell line resources highlighted above also provide preclinical efficacy data in the form of IC_{50} for each therapeutic-cancer cell line combination.[78] Within the context of cancer treatment, IC_{50} refers to the minimum dosage required to inhibit 50% of the cancer cells. While IC_{50} is an indicator of potential efficacy, the relationship between the IC_{50} value and drug approval is unclear. The IC_{50} value is an *in vitro* measurement, therefore translation into clinical efficacy is not guaranteed. Furthermore, it does not take into account the potential toxicity of the therapeutic being investigated.

Clinical outcomes

The focus of most drug efficacy models has been the preclinical IC_{50} measurement as a number of public databases, such as Cancer Cell Line Encyclopedia (CCLE) (Broad Institute, https://sites.broadinstitute.org/ccle/) and Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org/), provide those data in a centralized location. In the case of clinical outcomes, the data are less centralized and require a fair amount of curation. The major resource for clinical outcomes is ClinicalTrials.gov, a registry of clinical trials run by the US National Library of Medicine. However, the data provided require manual amending and curation.

In oncology a number of key clinical endpoints are used to assess clinical efficacy.

1. Objective response rate (ORR): the percentage of patients who respond to treatment in a defined manner, e.g., the tumor shrinks or disappears.
2. Progression-free survival: the median or mean period of time that each patient spends without the disease showing any progression or advancing further.
3. Overall survival: the median or mean period of time that each patient, who takes a particular treatment, survives post-treatment.

A particular challenge in applying the preclinical cell line approach to patient data is that there are comparatively fewer datasets where the -omic profiles, treatment, and response of the patients are all available.

Models of interest to predict efficacy

A number of models have been developed to predict the IC_{50} of drug-cell line combinations. In Table 2 we list a number of the

models that have emerged in the past few years as well as their Spearman correlation as an assessment metric. All the models follow the same core idea of having therapeutic and disease state representations with the goal of predicting the IC_{50} of a drug and cell line combination. The biggest differences are what the models use to represent the therapeutic as well as the disease state and the underlying architecture of the neural network.

Clinical efficacy modeling

Similar approaches have yet to emerge to predict clinical efficacy directly. The limitations described above regarding clinical data make adopting the preclinical framework challenging. Specifically, the biggest issue is the lack of large databases of patient outcomes with multiple treatment options.

To address this issue, we can create patient populations representations called virtual-cohorts based on: (1) cancer type, (2) stage, (3) demographic information, and (4) biomarkers. The response of these virtual cohorts to different therapies is then considered as independent data points, with a representative -omic profile for each cohort generated. In Figure 3C we show the results of a mixed model where we take the transcriptomic profiles of cancer patients and use an IC_{50} predictor[6] to model efficacy. Figure 3C is the same plot as in Figure 3B; however, rather than the TCGA subtype as the hue color, we show the predicted efficacy value for each patient and a representative therapeutic, in this case eribulin. As a result, we have a proxy for how predictive each of the treatments will be for each of the patients with their unique expression pattern.

We then integrate the predicted efficacy over the patients for each subtype and plot the expected efficacy of the therapeutic for each indication subtype, as shown in Figure 3D. It is important to highlight that we could have also found the predicted efficacy for cluster generated from the embedding itself. However, we chose to focus on the canonical subtypes to compare the predicted results with data in the literature. In this particular example, the results are consistent as eribulin has been found to be more effective against triple-negative "basal" breast cancer.[79] For a more systematic assessment of this approach, we can utilize a dataset of 194 therapeutics that have either been approved or rejected by the FDA for a set of 14 oncology subtypes, and we can assess how the predicted IC_{50} values correspond to their clinical potential.

Before establishing the efficacy of the predicted models, we should first set up a baseline of how predictive the real IC_{50} values are of eventual approval for 74 distinct therapeutics present in CCLE. In the top panel of Figures 4A and 4B we show the IC_{50} value distributions of drug-disease pairs collected from CCLE for both the approved and the rejected drugs. The results show that therapeutics with a low IC_{50} value against cancer cell lines have a higher historical approval rate than those with higher IC_{50} values. Notice, however, that a low IC_{50} is not a guarantee that the drug gains approval, as a number of low IC_{50} drug-disease pairs end up getting rejected. The IC_{50} is a measure of how effective the drug is at inhibiting a cancer model cell line. On its own, it gives no information on its ability to target healthy cells. Also it does not provide a sense for how it might behave within the context of the human body. However, it is quite clear that it has real predictive value.

In the middle panels of Figures 4A and 4B we show the results using the predicted values for IC_{50} for the various therapeutic agents against the CCLE cancer cell lines. The relationship between IC_{50} and approval is consistent with the real data.

In the bottom panels of Figures 4A and 4B we also show predicted IC_{50} values using the patient transcriptomic profiles collected from TCGA. A similar pattern holds with the drug-disease pairs that are expected to have low IC_{50} values, and have correspondingly higher rates of approval historically. The results are summarized in Figure 4C, where the IC_{50} distributions shown are binned and their historical approval rate is calculated. There is a consistent pattern between the real and the predicted IC_{50} values, with the high-efficacy models having an increased probability for approval.

TOXICITY

Any therapeutic that seeks to gain FDA approval must have an acceptable safety profile. Therefore, being able to predict the potential toxicity of a new therapeutic agent is just as important and assessing its efficacy.

Developing models to predict toxicity requires access to reliable large-scale data for assessment of various chemical agents. The US Tox21 program is an initiative that has developed a number of *in vitro* assays that utilize quantitative high-throughput screening to generate a large number of toxicity measurements for thousands of various chemical agents.[80] The Tox21 *in vitro* assays are reported to be as reliable as animal models in predicting human toxicity levels[81] and have clear utility in predicting adverse effects of a drug.[82] The massive Tox21 dataset has been used to develop multiple ML models for predicting toxicity as part of the Tox21 challenge.[83] One of the best performing models achieved an ROC-AUC of 0.88 on predicting Tox21 data.[84] The toxicity prediction can then be utilized by other higher-order models to assess the likelihood of approval for new possible therapeutics.

In Figure 4D we show the relationship between the predicted toxicity[84] and the approval rate. As was expected, the therapeutics with lower predicted toxicity have a higher historical approval rate. The relationship in itself is not surprising; however, it is worth noting that the toxicity value used is a purely predicted value using a model that only requires a representation of the therapeutic, in this case an SMILE structure.

PREDICTING LIKELIHOOD OF FDA APPROVAL FOR THERAPEUTICS

In Figure 4E we developed a simple random-forest classifier model to predict whether a drug gains approval for a specific indication. The models use only a few features, which are highlighted on the x axis: the clinical ORR, the predicted IC_{50}, and the predicted toxicity. We show the results of a 10-fold cross-validated model for each of the feature sets, as summarized in Table 3. The ORR is in itself predictive of approval (AUC = 0.83), but has a wide spread in the AUC between various folds. The inclusion of the predicted IC_{50} and the toxicity improves the predictions and creates more consistent predictions

CellPress
OPEN ACCESS

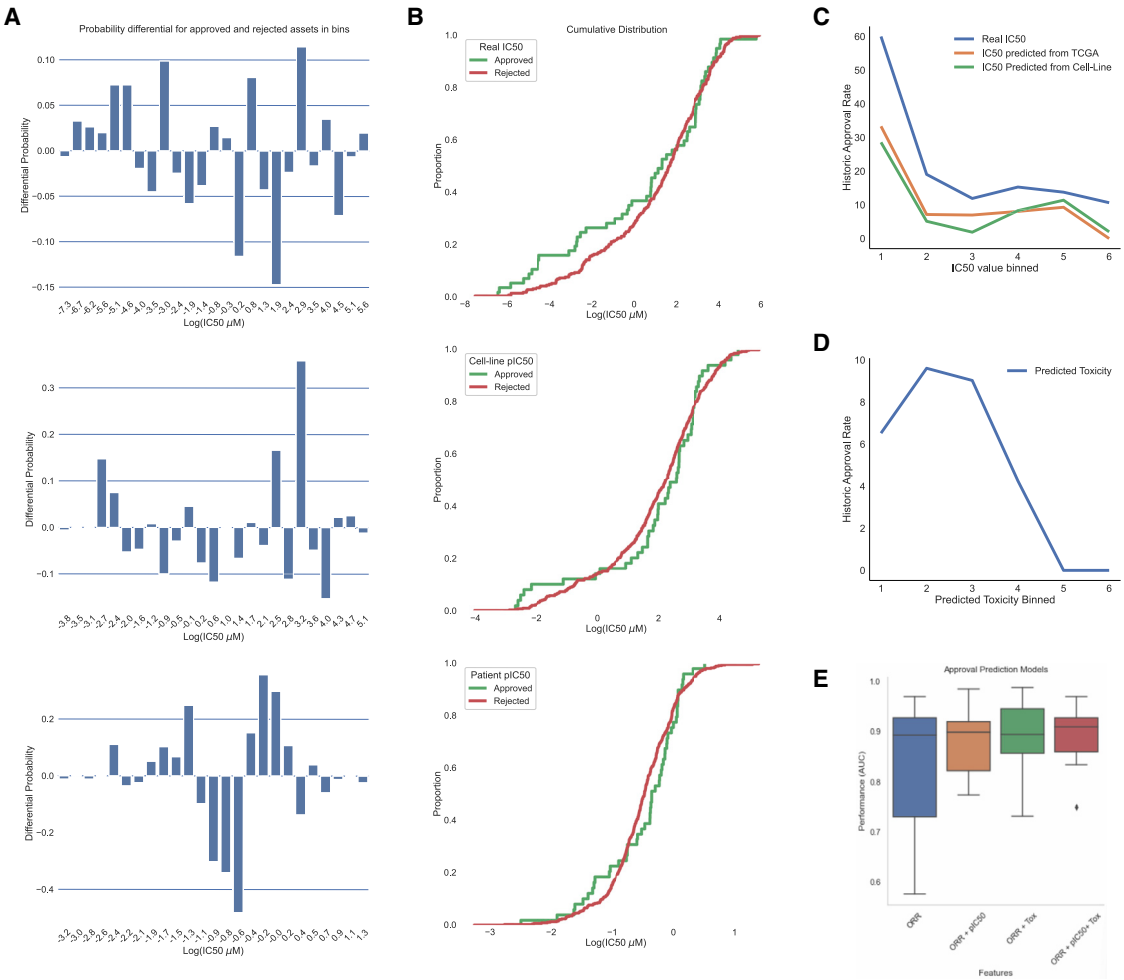

Figure 4. Exploring the relationship between drug-disease IC_{50} value and the likelihood of approval

(A) Probability distributions differential for the IC_{50} of a drug disease where the probability difference of the drug being approved or rejected is calculated for each bin. (Top) Real IC_{50} values collected from CCLE and averaged over the indication subtype that each cell line is associated with. (Middle) Predicted IC_{50} values using a previously published model[6] using the transcriptomic profiles of the CCLE cell lines. (Bottom) Predicted IC_{50} values using the same model but utilizing patient transcriptomic data collected from TCGA.

(B) Same as (A) but plotted as cumulative probability distributions.

(C) Calculated historical approval rates for different drug-disease combination of the probability distribution. Bins of low IC_{50} values, i.e., high efficacy, have a higher historical approval rate than those of high IC_{50}/low efficacy. The three lines correspond to the three data types: real IC_{50} (blue), cell line predicted IC_{50} (green), and patient predicted IC_{50} (orange).

(D) The historical approval rates of the binned toxicity for each of the drugs predicted using previously published models.[80] Lower predicted toxicity drugs have on average higher historical approval rates than high toxicity drugs.

(E) Simple random forest model predicting approval using only a few key features (objective response rate, predicted patient IC_{50}, and predicted toxicity). As can be seen, inclusion of predicted toxicity and IC_{50} improves the predictive capacity of the model. The mid-line represents the median, and the box-length shows the interquartile range (IQR) between the 25th and 75th quartiles of the data. The whiskers represent the range of the remaining data if they fall within 1.5x of the IQR range; otherwise the data points are plotted as outliers.

(AUC = 0.89), with a much tighter standard deviation of 0.06 compared with 0.13 for the ORR alone.

FUTURE DIRECTIONS: WHERE WE STAND AND WHERE WE ARE HEADING

In this perspective, we lay out the basic schema of the approach that many AI models take in the domain of drug discovery and design. We also review the fundamentals in terms of model types, data sources, and the potential insights each provide. Subsequently, we show the ability of these models to inform the likelihood of approval by utilizing the predicted efficacy and toxicity of a potential therapeutic.

Yet, there are a number of areas of active research that we did not touch upon so far. Within the domain of therapeutic representation most of the current work has focused on small-molecule therapies, as they are the most tractable. Predicting the efficacy and toxicity of higher-order therapeutics, such as

Table 3. AUC values for random forest models predicting the approval of drug-indication pairs using the features listed

Features	Average AUC (10-fold CV)	STD AUC (10-fold CV)
ORR	0.83	0.13
ORR + predicted toxicity	0.89	0.08
ORR + predicted IC_{50}	0.88	0.07
ORR + predicted toxicity + predicted efficacy	0.89	0.06

large proteins, mRNA therapies, and cell therapies, are lacking. These advanced therapeutic types and their associated representations are an area of active research and are expected to advance significantly in the near term.

For the representation of the disease state, we looked at the different -omic profiles as a way to capture the relevant information. While this approach is appropriate for diseases, such as oncology and autoimmune diseases, they are not directly transferable to bacterial or viral diseases. There, representation of the pathogen of interest would be more appropriate.

In terms of the model types we discuss, we look at supervised and unsupervised learning, but we did not delve into reinforcement learning (RL) (a form of ML wherein optimal strategies are found by defining an agent, an environment, and a cost function) or generative models.[6] In RL models the approach is quite different, as the researcher must *a priori* define a state space or "environment," an agent with well-defined actions within the environment, and a cost function to be optimized for a particular task. Moreover, these models can be combined with generative ones and efficacy predictors to develop novel therapeutics that are designed to target specific disease states.[6]

The use of MLA for the purposes of drug discovery, assessment, and design is still in its infancy. Despite recent advances, it is quite evident that the future will bring even more rapid and consequential applications of MLA in this field.

ACKNOWLEDGMENTS

V.G.G. is financially supported by the National Public Investment Program of the Ministry of Development and Investment/General Secretariat for Research and Technology, in the framework of the Flagship Initiatives to address SARS-CoV-2 (2020ΣΕ01300001) and "Application of artificial intelligence in drug development and new therapies" (2021NA11900006); the European Regional Development Fund of the European Union and Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-02939); the Hellenic Foundation for Research and Innovation (HFRI) grant nos. 775 and 3782; and NKUA SARG grant 70/3/8916.

DECLARATION OF INTERESTS

B.B. is an employee of Intelligencia Inc. G.L. is a co-founder of Intelligencia Inc. D.S. is a founder of Intelligencia Inc. A.T. and V.G.G. are advisors of Intelligencia Inc.

REFERENCES

1. Schuhmacher, A., Gatto, A., Kuss, M., Gassmann, O., and Hinder, M. (2021). Big Techs and startups in pharmaceutical R&D – a 2020 perspective on artificial intelligence. Drug Discov. Today *26*, 2226–2231. https://doi.org/10.1016/j.drudis.2021.04.028.
2. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., and Tekade, R.K. (2021). Artificial intelligence in drug discovery and development. Drug Discov. Today *26*, 80–93. https://doi.org/10.1016/j.drudis.2020.10.010.
3. Vougas, K., Sakellaropoulos, T., Kotsinas, A., Foukas, G.-R.P., Ntargaras, A., Koinis, F., Polyzos, A., Myrianthopoulos, V., Zhou, H., Narang, S., et al. (2019). Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining. Pharmacol. Ther. *203*, 107395. https://doi.org/10.1016/j.pharmthera.2019.107395.
4. Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A., Moss, T.J., Piha-Paul, S., Zhou, H., Kardala, E., et al. (2019). A deep learning framework for predicting response to therapy in cancer. Cell Rep. *29*, 3367–3373.e4. https://doi.org/10.1016/j.celrep.2019.11.017.
5. Kuenzi, B.M., Park, J., Fong, S.H., Sanchez, K.S., Lee, J., Kreisberg, J.F., Ma, J., and Ideker, T. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell *38*, 672–684.e6. https://doi.org/10.1016/j.ccell.2020.09.014.
6. Born, J., Manica, M., Oskooei, A., Cadow, J., Markert, G., and Rodríguez Martínez, M. (2021). PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. iScience *24*, 102269. https://doi.org/10.1016/j.isci.2021.102269.
7. Hazapi, O., Lagopati, N., Pezoulas, V.C., Papayiannis, G., Fotiadis, D.I., Skaltsas, D., Vergetis, V., Tsirigos, A., Stratis, I.G., and Yannacopoulos, A.N. (2022). Machine learning: a tool to shape the future of medicine. In Handbook of Machine Learning Applications for Genomics (Springer), pp. 177–218.
8. Solberg, H.E. (1978). Discriminant analysis. CRC Crit. Rev. Clin. Lab. Sci. *9*, 209–242. https://doi.org/10.3109/10408367809150920.
9. Ghojogh, B., Ghodsi, A., Karray, F., and Crowley, M. (2021). Uniform Manifold approximation and projection (UMAP) and its variants: tutorial and survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2109.02508.
10. GTEx Consortium; Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585. https://doi.org/10.1038/ng.2653.
11. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.
12. Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. *19*, A68–A77. https://doi.org/10.5114/wo.2014.47136.
13. ENCODE Project Consortium; and Pachter, L. (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science *306*, 636–640.
14. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612. https://doi.org/10.1093/nar/gkaa1074.
15. Schmidt, T., Samaras, P., Frejno, M., Gessulat, S., Barnert, M., Kienegger, H., Krcmar, H., Schlegl, J., Ehrlich, H.-C., Aiche, S., et al. (2018). ProteomicsDB. Nucleic Acids Res. *46*, D1271–D1281. https://doi.org/10.1093/nar/gkx1029.
16. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457–D462. https://doi.org/10.1093/nar/gkv1070.
17. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). The

reactome pathway knowledgebase 2022. Nucleic Acids Res. *50*, D687–D692. https://doi.org/10.1093/nar/gkab1028.

18. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2021). PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. *49*, D1388–D1395. https://doi.org/10.1093/nar/gkaa971.
19. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.
20. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45*, D945–D954. https://doi.org/10.1093/nar/gkw1074.
21. Nayak, J., Naik, B., and Behera, P.D.H. (2015). Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014, pp. 133–149. https://doi.org/10.1007/978-81-322-2208-8_14.
22. Carreira-Perpiñán, M.Á. (2015). A review of mean-shift algorithms for clustering. Preprint at arXiv. https://doi.org/10.48550/arXiv.1503.00687.
23. Lakshmi, T.M., Sahana, R.J., and Venkatesan, V.P. (2018). Review on density based clustering algorithms for big data. IJDMA *7*, 13–20.
24. Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H.J., and Zacharias, H.U. (2020). Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools. Biochim. Biophys. Acta. Gene Regul. Mech. *1863*, 194418. https://doi.org/10.1016/j.bbagrm.2019.194418.
25. Schmidhuber, J. (2015). Deep learning in neural networks: an overview. Neural Network. *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.
26. Lipinski, C.A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today Technol. *1*, 337–341. https://doi.org/10.1016/j.ddtec.2004.11.007.
27. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. *28*, 31–36.
28. Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach. Learn, Sci. Technol. *1*, 045024. https://doi.org/10.1088/2632-2153/aba947.
29. David, L., Thakkar, A., Mercado, R., and Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminf. *12*, 56. https://doi.org/10.1186/s13321-020-00460-5.
30. Morgan, H.L. (1965). The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. J. Chem. Doc. *5*, 107–113. https://doi.org/10.1021/c160017a018.
31. Capecchi, A., Probst, D., and Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J. Cheminf. *12*, 43. https://doi.org/10.1186/s13321-020-00445-4.
32. Jastrzębski, S., Leśniak, D., and Czarnecki, W. (2016). Learning to SMILE(S).
33. Trinajstic, N. (1992). Chemical Graph Theory, 2nd Edition (CRC Press). https://doi.org/10.1201/9781315139111.
34. Nguyen, T., Nguyen, G.T.T., Nguyen, T., and Le, D.H. (2022). Graph convolutional networks for drug response prediction. IEEE ACM Trans. Comput. Biol. Bioinf *19*, 146–154. https://doi.org/10.1109/tcbb.2021.3060430.
35. Liu, Q., Hu, Z., Jiang, R., and Zhou, M. (2020). DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics *36*, i911–i918. https://doi.org/10.1093/bioinformatics/btaa822.
36. Feng, R., Xie, Y., Lai, M., Chen, D.Z., Cao, J., and Wu, J. (2021). AGMI: attention-guided multi-omics integration for drug response prediction with graph neural networks, pp. 1295–1298.
37. Yang, K.K., Wu, Z., Bedbrook, C.N., and Arnold, F.H. (2018). Learned protein embeddings for machine learning. Bioinformatics *34*, 2642–2648. https://doi.org/10.1093/bioinformatics/bty178.
38. Mikolov, T., Chen, K., Corrado, G.s., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of Workshop at ICLR 2013.
39. Quoc, L., and Tomas, M. (2014). Distributed Representations of Sentences and Documents (PMLR).
40. Asgari, E., and Mofrad, M.R.K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One *10*, e0141287. https://doi.org/10.1371/journal.pone.0141287.
41. Kimothi, D., Soni, A., Biyani, P., and Hogan, J.M. (2016). Distributed representations for biological sequence analysis. Preprint at arXiv. https://doi.org/10.48550/arXiv.1608.05949.
42. Ng, P. (2017). dna2vec: consistent vector representations of variable-length k-mers. Preprint at arXiv. https://doi.org/10.48550/arXiv.1701.06279.
43. Mazzaferro, C. (2017). Predicting protein binding affinity with word embeddings and recurrent neural networks. Preprint at bioRxiv. https://doi.org/10.1101/128223.
44. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y.S. (2019). Evaluating protein transfer learning with TAPE. Adv. Neural Inf. Process. Syst. *32*, 9689–9701.
45. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). In Recurrent Neural Network Based Language Model, *3* (Makuhari), pp. 1045–1048.
46. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
47. Minhas, F.u.A.A., Geiss, B.J., and Ben-Hur, A. (2014). PAIRpred: partner-specific prediction of interacting residues from sequence and structure. Proteins *82*, 1142–1155. https://doi.org/10.1002/prot.24479.
48. Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. Adv. Neural Inf. Process. Syst. *30*.
49. Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nat. Commun. *12*, 3168. https://doi.org/10.1038/s41467-021-23303-9.
50. Wallach, I., Dzamba, M., and Heifets, A. (2015). AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. Preprint at arXiv. https://doi.org/10.48550/arXiv.1510.02855.
51. Taylor, R.D., Jewsbury, P.J., and Essex, J.W. (2002). A review of protein-small molecule docking methods. J. Comput. Aided Mol. Des. *16*, 151–166. https://doi.org/10.1023/A:1020155510718.
52. Ruppert, J., Welch, W., and Jain, A.N. (1997). Automatic identification and representation of protein binding sites for molecular docking. Protein Sci. *6*, 524–533. https://doi.org/10.1002/pro.5560060302.
53. Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V., Oprea, T.I., Baskin, I.I., Varnek, A., Roitberg, A., et al. (2020). QSAR without borders. Chem. Soc. Rev. *49*, 3525–3564. https://doi.org/10.1039/D0CS00098A.
54. Ding, X., Tsang, S.-Y., Ng, S.-K., and Xue, H. (2014). Application of machine learning to development of copy number variation-based prediction of cancer risk. Genomics Insights *7*, 1–11. https://doi.org/10.4137/GEI.S15002.
55. Li, M., Wang, Y., Zheng, R., Shi, X., Li, Y., Wu, F.X., and Wang, J. (2021). DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines. IEEE ACM Trans. Comput. Biol. Bioinf *18*, 575–582. https://doi.org/10.1109/tcbb.2019.2919581.
56. Chang, Y., Park, H., Yang, H.-J., Lee, S., Lee, K.-Y., Kim, T.S., Jung, J., and Shin, J.-M. (2018). Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. Sci. Rep. *8*, 8857. https://doi.org/10.1038/s41598-018-27214-6.

57. Liu, P., Li, H., Li, S., and Leung, K.-S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinf. *20*, 408. https://doi.org/10.1186/s12859-019-2910-6.

58. Wang, F., Bai, X., Wang, Y., Jiang, Y., Ai, B., Zhang, Y., Liu, Y., Xu, M., Wang, Q., Han, X., et al. (2021). ATACdb: a comprehensive human chromatin accessibility database. Nucleic Acids Res. *49*, D55–D64. https://doi.org/10.1093/nar/gkaa943.

59. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

60. Sherman, M.A., Yaari, A.U., Priebe, O., Dietlein, F., Loh, P.R., and Berger, B. (2022). Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. Nat. Biotechnol. *40*, 1634–1643. https://doi.org/10.1038/s41587-022-01353-8.

61. Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., et al. (2020). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. *48*, D489–D497. https://doi.org/10.1093/nar/gkz946.

62. Türei, D., Korcsmáros, T., and Saez-Rodriguez, J. (2016). OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat. Methods *13*, 966–967. https://doi.org/10.1038/nmeth.4077.

63. Oskooei, A., Manica, M., Mathis, R., and Martínez, M.R. (2019). Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. Sci. Rep. *9*, 15918–16013.

64. Chen, J., and Zhang, L. (2021). A survey and systematic assessment of computational methods for drug response prediction. Briefings Bioinf. *22*, 232–246. https://doi.org/10.1093/bib/bbz164.

65. Dugourd, A., Kuppe, C., Sciacovelli, M., Gjerga, E., Gabor, A., Emdal, K.B., Vieira, V., Bekker-Jensen, D.B., Kranz, J., Bindels, E.M.J., et al. (2021). Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. Mol. Syst. Biol. *17*, e9730. https://doi.org/10.15252/msb.20209730.

66. Simidjievski, N., Bodnar, C., Tariq, I., Scherer, P., Andres Terre, H., Shams, Z., Jamnik, M., and Liò, P. (2019). Variational autoencoders for cancer data integration: design principles and computational practice. Front. Genet. *10*, 1205. https://doi.org/10.3389/fgene.2019.01205.

67. Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P., and Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst. *6*, 484–495.e5. https://doi.org/10.1016/j.cels.2018.03.001.

68. Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proc. IEEE *104*, 11–33. https://doi.org/10.1109/JPROC.2015.2483592.

69. Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Bender, A., Hoyt, C.T., and Hamilton, W.L. (2022). A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. Briefings Bioinf. *23*, bbac404. https://doi.org/10.1093/bib/bbac404.

70. Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Hoyt, C.T., and Hamilton, W.L. (2022). Understanding the performance of knowledge graph embeddings in drug discovery. Artificial Intelligence in the Life Sciences *2*, 100036. https://doi.org/10.1016/j.ailsci.2022.100036.

71. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction (PMLR), pp. 2071–2080.

72. Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6575.

73. Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. (2019). Rotate: knowledge graph embedding by relational rotation in complex space. Preprint at arXiv. https://doi.org/10.48550/arXiv.1902.10197.

74. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. *26*.

75. Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI conference on artificial intelligence, p. 1.

76. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife *6*, e26726. https://doi.org/10.7554/eLife.26726.

77. Walsh, B., Mohamed, S.K., and Nováček, V. (2020). Biokg: a knowledge graph for relational learning on biological data, pp. 3173–3180.

78. Vis, D.J., Bombardelli, L., Lightfoot, H., Iorio, F., Garnett, M.J., and Wessels, L.F. (2016). Multilevel models improve precision and speed of IC50 estimates. Pharmacogenomics *17*, 691–700. https://doi.org/10.2217/pgs.16.15.

79. Pizzuti, L., Krasniqi, E., Barchiesi, G., Mazzotta, M., Barba, M., Amodio, A., Massimiani, G., Pelle, F., Kayal, R., Vizza, E., et al. (2019). Eribulin in triple negative metastatic breast cancer: critic interpretation of current evidence and projection for future scenarios. J. Cancer *10*, 5903–5914. https://doi.org/10.7150/jca.35109.

80. Huang, R. (2016). A quantitative high-throughput screening data analysis pipeline for activity profiling. Methods Mol. Biol. *1473*, 111–122. https://doi.org/10.1007/978-1-4939-6346-1_12.

81. Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S.A., Attene-Ramos, M., Zhao, T., Austin, C.P., and Simeonov, A. (2016). Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. Nat. Commun. *7*, 10425. https://doi.org/10.1038/ncomms10425.

82. Huang, R., Xia, M., Sakamuru, S., Zhao, J., Lynch, C., Zhao, T., Zhu, H., Austin, C.P., and Simeonov, A. (2018). Expanding biological space coverage enhances the prediction of drug adverse effects in human using in vitro activity profiles. Sci. Rep. *8*, 3783. https://doi.org/10.1038/s41598-018-22046-w.

83. Huang, R., Xia, M., Nguyen, D.-T., Zhao, T., Sakamuru, S., Zhao, J., Shahane, S.A., Rossoshek, A., and Simeonov, A. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front. Environ. Sci. *3*. https://doi.org/10.3389/fenvs.2015.00085.

84. Markert, G., Born, J., Manica, M., Schneider, G., and Rodriguez Martinez, M. (2020). Chemical representation learning for toxicity prediction.