Expert Opinion on Drug Discovery

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iedc20

The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction

David Z Huang, J. Christian Baber & Sogole Sami Bahmanyar

To cite this article: David Z Huang, J. Christian Baber & Sogole Sami Bahmanyar (2021) The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction, Expert Opinion on Drug Discovery, 16:9, 1045-1056, DOI: 10.1080/17460441.2021.1901685

To link to this article: https://doi.org/10.1080/17460441.2021.1901685

Published online: 19 Mar 2021.

Submit your article to this journal

Article views: 525

View related articles

View Crossmark data

Citing articles: 1 View citing articles

EXPERT OPINION ON DRUG DISCOVERY
2021, VOL. 16, NO. 9, 1045–1056
https://doi.org/10.1080/17460441.2021.1901685

REVIEW

The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction

David Z Huang[a], J. Christian Baber[b] and Sogole Sami Bahmanyar[c]

[a]Harvard College, Harvard University, Cambridge, USA; [b]Scientific Informatics, Global Head of Scientific Informatics, Scientific Informatics, Takeda Pharmaceuticals, Cambridge, MA, USA; [c]Computational Chemistry, Director of Computational Sciences, Computational Chemistry, Takeda Pharmaceuticals, San Diego, USA

ABSTRACT

Introduction: Artificial intelligence (AI) has seen a massive resurgence in recent years with wide successes in computer vision, natural language processing, and games. The similar creation of robust and accurate AI models for ADME/Tox endpoint and activity prediction would be revolutionary to drug discovery pipelines. There have been numerous demonstrations of successful applications, but a key challenge remains: how generalizable are these predictive models?

Areas covered: The authors present a summary of current promising components of AI models in the context of early drug discovery where ADME/Tox endpoint and activity prediction is the main driver of the iterative drug design process. Following that is a review of applicability domains and dataset construction considerations which determine generalizability bottlenecks for AI deployment. Further reviewed is the role of promising learning frameworks – multitask, transfer, and meta learning – which leverage auxiliary data to overcome issues of generalizability.

Expert opinion: The authors conclude that the most promising direction toward integrating reliable and informative AI models into the drug discovery pipeline is a conjunction of learned feature representations, deep learning, and novel learning frameworks. Such a solution would address the sparse and incomplete datasets that are available for key endpoints related to drug discovery.

ARTICLE HISTORY
Received 11 January 2021
Accepted 8 March 2021

KEYWORDS
Artificial intelligence (AI); quantitative structure–activity relationship (QSAR); activity prediction; ADME/Tox endpoints; generalizability; deep learning (DL); applicability domain (AD); transfer learning; meta learning; multitask learning

1. Introduction

Artificial intelligence (AI) techniques and deep learning (DL) techniques in particular have delivered incredible results in domains as diverse as computer vision [1], natural language processing [2], and Go [3]. To clarify terms: AI is an umbrella term for all reasoning computers; machine learning (ML) is a subset of AI where models are constructed from experience and data; DL [4,5] is a subset of ML, where models are constructed of complex webs of artificial neurons, which are nodes performing simple computations.

AI's broad success has prompted much hope that it could be useful for drug discovery and repurposing [6–9], and funding for its application is expected to reach 20 billion dollars by 2024 [10,11]. Indeed, recent landmark case studies have demonstrated the effectiveness of AI approaches in this domain. These include the discovery of novel antibiotics [12] and AlphaFold2's successful prediction of protein structure from sequences [13], where they reached a score of almost 90 on the Global Distance Test, roughly equivalent to experimental accuracy [13], further improving on their first AI-based model [14]. Although large AI models have been traditionally cost and resource prohibitive to train, innovations in computer technologies have been breaking those barriers.

One of the key areas in drug discovery benefitting from AI is ADME/Tox endpoint and activities prediction where AI models are used as quantitative structure–activity relationship (QSAR) models for the prediction of a variety of properties, ranging from more straightforward physicochemical predictions to the more complex pharmacokinetic, pharmacokinetic, and toxicological properties [6,15–18]. Important pharmacokinetic (PK) endpoints include clearance, permeability, and stability; important pharmacodynamic (PD) endpoint include drug-target specificity and selectivity; important toxicological endpoints include cytochrome P450 induction and hERG inhibition.

The availability of predictive models as part of the design cycle is extremely valuable as it can allow for chemists and biologists to effectively triage molecules and select a drug candidate with the highest likelihood of success in the clinic. These *in silico* tools can be time and cost saving and decrease the expense and high attrition of introducing new medicines to the patient [9].

Many studies have been undertaken to evaluate the new emerging AI models against traditional machine learning and cheminformatics models. In the first of these endeavors, Merck hosted a Kaggle competition [19], challenging contestants to build models for 15 different and diverse QSAR datasets. The winning entry used an ensemble of methods with the primary predictor being a deep neural network (DNN). Analyzing these results in 2015, Ma et al. [20] found that the simple application of DNNs routinely outperformed random forests (RFs). In the analysis of another of these endeavors

CONTACT David Z Huang dhuang@college.harvard.edu Harvard College, Harvard University, Cambridge, MA, USA02138

Article highlights

- Harnessing computational power to make predictions based on data, AI is an effective technique and has proven itself in other domains and is poised to do the same to drug discovery.
- AI models that make robust *in silico* predictions for activities and ADME/Tox endpoints can help make the drug candidate selection more efficient and cost effective, delivering safe and efficacious medicines for the patient.
- The main challenge to creating accurate and applicable AI models is that the available experimental data is heterogenous, noisy, and sparse, so appropriate data curation and data collection is of the utmost importance.
- Ultimately, the ability of AI models to become more generalizable to novel situations is the only solution to keep up with the fast-paced and dynamic nature of drug discovery.
- The most promising, generalizable AI models utilize deep learning on learned feature representations, integrated under learning frameworks which allow models to aggregate information from similar but distinct ADME/Tox endpoint and activity prediction tasks.

This box summarizes key points contained in the article.

2018, Mayr et al. [21] undertook a detailed comparison of machine learning (ML) models for drug target prediction using the ChEMBL database [22], similarly finding that DNNs outperform RFs, as well as support vector machines (SVMs), K-nearest-neighbors (KNN), Naïve Bayes (NB) and a similarity ensemble approach (SEA).

However, the aforementioned AI models have struggled to make meaningful predictions beyond compounds with molecular representations that are highly similar to their data [23,24]. In a recent study, Su et al. [25] evaluated the performance of ML algorithms for protein–ligand interaction scoring. They found that the scoring power gains of their best algorithm, RFs, were heavily dependent on the similarity of the train and test set. Similarly, Liu et al. [26] found that for ADME/Tox endpoint prediction of *in vivo* acute chemical toxicity and for the aforementioned Merck Kaggle datasets, the performance of AI algorithms–DNNs, RFs, and variable nearest neighbors (v-NN)–was highly dependent on the similarity between the training set and the test set.

This limitation means that AI models will not be able to make accurate predictions for the novel compounds that we want it to elucidate. Indeed, the coming challenge lies in improving the ability of AI to generalize from very little data to a broader chemical space. The development of such a class of models will make the large-scale application of AI practical and will be revolutionary, reducing drug candidate attrition rates and research costs [7,10,16,27,28].

To some degree, this is an inherent, intractable problem; after all, AI models can only make predictions based on the available data [29]. And, drug discovery data is both highly sparse – with few experimental data points in the vast chemical space of possibilities–and often noisy – depending on assay accuracy and sensitivity [30]. But this challenge is not unsurmountable. For example, in a study on 13,000 compounds against a 159-assay kinase panel with only experimental data for 5% of possible values, researchers used data imputation to accurately fill in the remaining 95% of values [31].

This is a problem also present in the broader AI field, where there exists suites of tools and case studies of successful modeling of data scarce environment [32–34], which will be discussed in more detail later. So, there is still great potential for the development of AI models for property prediction, and indeed there has been a rapid growth in novel model architectures proposed and case study applications.

2. Model architecture

The state of AI molecular property prediction is quickly evolving with an explosion in the number of new models and permutations thereof. Despite the variety in new models, each one has the same two-part structure: (1) a molecular representation which encodes a compound in a computer-readable format and (2) an AI algorithm which makes a prediction based on a molecular representation. These choices are the defining factor in how accurate and how generalizable an AI model is.

Before proceeding, it should be noted that in the current state of the field it is tricky to draw broad conclusions from narrow studies. Molecular representations and AI algorithms work better or worse depending on the characteristics of the dataset they are trained and evaluated on, and subtle differences in architectures and parameter setups can make a world of difference.

2.1. Molecular representation

There is a broad range of molecular representations [35] to choose from which are summarized in Table 1.

Table 1. The most commonly used and promising molecular representations for small molecules.

Molecular representation name	Description
Molecular descriptors	A vector of numbers which encode certain computed or experimental physicochemical, structural, topological, and/or electronic properties. The most commonly used and validated class of molecular descriptors are 2D descriptors [36].
Fingerprints	A specific, complex type of molecular descriptor which encodes a compound in its structural and functional patterns, typically as a binary string where each bit represents the presence or absence of some pattern. Commonly used fingerprints include structural keys [37] and circular fingerprints [38].
Molecular graph	A mathematical graph where vertices are atoms and edges are bonds or distances between atoms. Typically, additional features such as bond type, atom hybridization, or charge are used to describe the vertices and edges.
Graph-theoretical Matrix	A matrix which encodes the molecular graph.
Simplified Molecular Line Entry System (SMILES)	A string of characters linearly representing the molecular graph of a compound encoding atom and bond information.

In the early history of QSAR modeling, molecular descriptors were hand-tailored to specific reaction frameworks of very small datasets. If a researcher knows that a particular endpoint depends on one particular descriptor, then they can just use that descriptor; for example, Fieser et al. [39] could directly graph the relationship between antimalarial potency of naphthoquinones and ether-water distribution coefficients. However, in most cases, these relationships are more subtle, encompassing a combination of factors. So as researchers began to work with larger and more diverse datasets, the need for general-purpose descriptors, which model all potential sources of variation, became apparent [40] and through the years many classes of descriptors have been developed and tested. A complete survey of thousands of varying molecular descriptors can be found in a listing by Todeschini and Consonni [41].

First, descriptors should be carefully chosen and pruned to avoid overfitting, which is where a model memorizes spurious correlations in the data which have no actual predictive value. This step is often done automatically, but it must be done very carefully in order to avoid leaking data and thus invalidating the results of a model built using the chosen data. The Topliss–Costello rule-of-thumb recommends that for regression, the ratio between training examples and descriptors should be no less than 5 to 1 [42].

Second, descriptors should be as parametrically diverse as necessary. Different sets of descriptors cover different segments of physicochemical space and some more so than others [43]. Thus, it must be insured that a set of descriptors models all relevant interactions for some property, and that the descriptors are able to extend and generalize outside of the narrow training set provided. For example, in virtual screening, a model with diverse descriptors will generate more useful and diverse hit lists [44].

Third, descriptors should be as orthogonal as possible to each other in the descriptor space. This allows for greater interpretability of the model and also avoids incorporating redundant and distracting information into a model [45].

The difficulty in choosing the right descriptors lies in large part in the conflict between the first and second point. Especially when predicting on small datasets with unknown or ill-defined data frameworks, it is hard to find a reasonable set of descriptors (to avoid overfitting) while still encompassing all the requisite information for the model to make an accurate prediction. The broader point is that in this scheme – choosing and some set of molecular descriptors – researchers are required to decide which quantities are relevant and which are not. The model's predictive power hinges on that, but it is a difficult task.

The problem of selecting the right molecular descriptors is a feature engineering problem. Recent models attempt to overcome the feature selection property by directly learning a feature representation from the compound's structure, which is typically represented as a molecular graph, a graph-theoretical matrix, or as a SMILES [46]. Using learned features instead of engineered features has been a fundamental theme in the recent rise of artificial intelligence. For example, in computer vision the grand breakthrough was ImageNet, using DL and numerous convolutional layers instead of hand-engineered features such as edge detection or ridge detection [1]. Similarly, in Go with DeepMind's AlphaGo, the AI did not use any preprogrammed move sequences [3].

Nevertheless, the learned featured representation approach still faces difficulties. A key limitation is that a model needs a lot of data to be able to learn its own features without overfitting. Ponzoni et al. [47] found that on blood–brain barrier (BBB), and human intestinal absorption datasets on the order of 100 compounds each, their best feature selection and feature learning models performed roughly equivalently. In a study of many diverse and often small prediction tasks on ChEMBL, Mayr et al. [21] found that models operating on precomputed descriptors performed better than models operating directly on molecular graphs. However, in follow-up studies by Yang et al. [48] and Wu et al. [49], they described new methods of feature learning directly on molecular graphs and achieved results better than had been originally reported by Mayr et al. [21]. Another key limitation is that a purely learned features molecular representation does not include useful experimental information that may have been collected already. For example, when modeling cellular inhibition, a purely learned feature representation wouldn't include important available experimental protein-ligand activity data. These are just a few illustrative examples of learned feature representations, and there is a great more variety in model architectures which will be surveyed later in conjunction with a discussion of AI algorithms.

2.2. AI Algorithms

The application of AI to ADME/Tox and activity prediction is rapidly evolving with new models being proposed at a fast rate. An overview is provided in Table 2.

Similar to molecular representations, AI algorithms can be broadly classified as using engineered features or learned features. SVMs, kNNs, RFs, linear regression, and MLP models fall into the former category; CNNs, RNNs, and MPNNs, and transformers fall into the latter category. The division is not completely clear cut; however, many learned features models do also incorporate molecular descriptors as their inputs.

Out of the feature engineered models, MLP has been shown to at least match and even frequently outperform SVMs, kNNs, RFs, and linear regression models on datasets as diverse as solubility, cell growth inhibition, logD, and CLINT [20,50–52]. While both the Ma et al. [20] and Korotcov et al. [50] studies agreed that MLPs outperformed all other models, Korotcov et al. found that DNNs performed worse on the hERG endpoint, while Ma et al. found that DNNs performed markedly better. Although their model architectures are slightly different, those differences will only provide modest gains, not significant changes in performance. Indeed, the substantial difference is that Korotcov et al. used approximately 500 compounds in their training set while Ma et al. used approximately 50,000 compounds. This reflects an overarching theme in AI: a feature-learning model performs much better than a feature-engineered model on large, complex datasets [53].

Table 2. The most commonly used and promising AI algorithms.

AI algorithm name	Description
Support vector machines (SVMs)	Classifies data into categories by first mapping it into a higher-dimensional space (to have nonlinearity) and then finding a hyperplane to separate the data into categories.
k-Nearest neighbors (kNNs)	Performs classification or regression for a new compound using the property values of the *k* most similar compounds to it
Decision trees (DTs)	A decision tree is where each node of the tree is a yes or no test, leading to a final prediction of the leaf nodes. Both random forests (RFs) and gradient boosting methods (GBMs) aggregate the prediction of multiple decision trees, with RFs creating DTs in parallel and GBMs iteratively creates DTs to supplant each other's weaknesses.
Linear regression	Models the relationship between explanatory variables and the response with a linear equation, found from the data. There are many extensions of linear regression. If using multiple explanatory variables, this is called Multiple Linear Regression. If combined with principal component analysis, it is called partial least squares regression. A logistic function can be applied to the output of linear regression for binary classification tasks. An L2-norm penalty and the kernel trick can be applied to make it possible to model nonlinear functions.
Multilayer perceptron (MLP)	This is a multilayered, feed-forward, deep neural network, most commonly with stacks of fully connected layers. MLPs are a type of deep neural network (DNN), and most of the studies on DNNs have been done specifically on MLPs.
Convolutional neural networks (CNNs)	CNNs are like DNNs but they apply the same function repeatedly to subcomponents of the data. This has the advantage of imposing regularization and imposing shift invariance.
Recurrent neural networks (RNNs)	RNNs process a sequence of items as input (i.e. a SMILES string of character) which are then used to predict the value. Common RNNs include Gated Recurrent Units and Long Short-Term Memory.
Message passing neural networks (MPNNs)	A MPNN is a network framework formalized by Gilmer et al. [54], where a neural network is fed a molecular graph as input and processes the graph by first continually updating the hidden state of each atom with its neighbors and then finally reading out a prediction using another model.
Transformers	A DL network architecture built on Self-Attention layers [59], which process sequential data (i.e. SMILES) very well.

A large class of learned features models are message passing neural networks (MPNNs) which directly operate on molecular graphs. This framework was first formalized by Gilmer et al. [54] and was an extension of other prior successful studies such as gated graph neural networks [55] and deep tensor neural networks [56]. In large-scale follow-up studies, MPNNs were shown to broadly outperform SVMs, RFs, kNNs, Linear Regression, and MLP, but there are certain situations where using MPNNs is not advantageous [49]. In particular, Yang et al. [48] concluded that they underperform in three situations: when other models incorporate 3D information, when the dataset is small, and when the classes are imbalanced. Indeed, it was that MPNN, which led finally to the discovery of a novel antibiotic [12].

Another class of learned features model uses sequential data as its molecular representation. Most commonly the molecular representation is SMILES and the AI algorithms are either CNNs, RNNs, or Transformers. It should be noted that these AI Algorithms can be used and often are used in conjunction with other molecular representations. Chakravarti et al. [57] found success using LSTMs (a type of RNN) on SMILES in modeling Ames mutagenicity, inhibition of P. falciparum Dd2 and inhibition of Hepatitis C Virus. They found that the RNN-based approach performed better in generalizing to dissimilar test set compounds versus fragment-based models [57]. Lusci et al. [58] found success in flattening molecular graphs into sequential data and then using an ensemble of RNNs to predict aqueous solubility, achieving state-of-the-art results [58].

Transformers [59], which have been especially successful in the domain of natural language processing with OpenAI's GPT-3 [2], are now being applied toward ADME/Tox endpoint and activity prediction as well. Particularly intriguing is the fact that GPT-3 worked as a few-shot model, which is where an AI algorithm only needs to see a few examples before it can make quality predictions – ideal for improving generalizability for endpoint and activity prediction. Furthermore, the fact that SMILES is encoded as text data prompted the natural application of SMILES to endpoint and activity prediction [60]. In recent studies, Schwaller et al. [61] found that their transformer-based model outperformed all other methods in predicting the products of organic synthesis and Nayak et al. [62] found that their transformer-based model performed better than MPNNs for ADME/Tox endpoint prediction.

These AI algorithms often are developed under transfer learning, multitask learning, and meta learning frameworks, which have been shown to greatly increase a model's predictive power. These frameworks will be surveyed together in a later section after a discussion on data quality and applicability domains, which provide a lens for investigating those frameworks.

3. Data quality and applicability domain

3.1. Applicability domains

The applicability domain (AD) of a model is defined as the response and descriptor space in which the model can be legitimately applied to make a prediction [63]. The domain of applicability is an especially relevant concern as the drug discovery space is expanding beyond small molecules to address the more challenging and novel target space with new modalities [64–66]. The concept of an AD was extended by Hanser et al. [67] to the decision domain (DD) of a model shown in Figure 1 which is a hierarchy defining the space where a model can confidently make predictions in three stages: (1) applicability, (2) reliability, and (3) decidability. Applicability measures whether or not the prediction the model is being asked to make is its intended use case. Reliability measures whether or not the model has been given enough information in its training set to make an informed prediction. Decidability measures how practically useful a decision; for example, a final 50–50 prediction is not very useful, and neither would an uncertain prediction with high error bars. Importantly, all three stages must hold for a model to be able to be confidently used to make decisions. Although not explicitly stated, most AD methods fall into this framework, providing a useful mental model for analyzing model generalizability.

Applicability depends on both the data available and the molecular representation. Quantitatively, a range of molecular descriptors where the model is applicable can be defined by using a convex hull or a range box on the descriptor space [68]. However, qualitative methods to assess applicability are just as important. Important considerations are included but are not limited to the following as a complete characterization of applicability can only be done in coordination with domain experts: the query compound must be in the same chemical class as the available data; and the desired predicted endpoint must be the same as the data's, i.e. even for the target, IC_{50} cannot be predicted solely with percent inhibition data.

Reliability also depends on both the data and the choice of molecular representation. For a given query compound, this is found by comparing its molecular representation with that of the data, seeing if there are similar compounds which would be the supporting information for the model to make a reliable prediction [69]. This is also termed 'novelty detection', as in that novel, unseen compounds will have unreliable predictions. A study by Carrió et al. found that good measures of this are the distance between the query compound and the centroid of the training set, and the distance between the query compound and the most similar compound in the training set [70].

Decidability depends on the complete AI model and its prediction. While reliability assessed the quality of the supporting information, decidability assesses the quality of the model for the specific query compound. Many models have natural metrics for decidability: with an ensemble of multiple models, we can examine the amount of agreement or disagreement within the ensemble and with many classification models there is either a class probability outputted or an internal regression model which can be interpreted as a model's decidability for a query compound [71]. If a model does not have such natural metrics, Carrió et al. gives the following criteria for decidability: distance between the predicted property value and the closest property value in the training set; and the standard deviation error of the predictions of the top 5% most similar compounds in the training set [70]. There are further uncertainty quantification models, where another model is used to assess the decidability of the primary model [18]. This could be another AI model [72] or a statistical model in the case of conformal prediction [73]. The specific choice of a decidability metric varies depending on the underlying model and data.

The task of increasing a model's generalizability is to increase its applicability domain. Thus, a model's generalizability is limited by its three components: data, molecular representation, and the AI algorithm.

3.2. Data quality

For practical model-building, the data component is the fundamental limitation, because computation has become widely, and cheaply available so even large and complex AI algorithms and molecular representations can be easily tested and implemented. However, the reality of the drug discovery process complicates the data sets that are generated. Every experimental datapoint has to be measured by biologists and chemists which requires time and resources. Given the nature of early drug discovery projects, data are generated for smaller subsets of molecules – in the range of hundreds to a couple of thousands – which may be limited to target-specific chemotypes. In addition, assay conditions and read-outs change over time for specific endpoints. Furthermore, there are more data available for high-throughput primary assay endpoints and less data for the more resource-intensive PK/PD and tox endpoints that are more important in the later stages of drug candidate selection.

Data availability in drug discovery can be contrasted with other domains where AI has been successfully applied: in computer vision ImageNet has 1.2 million data points for image recognition; and in natural language processing for GPT3, researchers used 500 billion tokens of text data. This quantity of data is nowhere near available for any single endpoint, and only when considering the entirety of all bioactivities in PubChem do we get a comparable approximate of 270 million bioactivities [74]. Although data augmentation [75] – whereby the original dataset is enhanced by creating additional samples for the model to learn, for example, by sampling different SMILES [76] or different conformations for the same compound [77] – is a powerful technique, it does not change the fact there are still only experimental data for a limited set of compounds, whereby other unique chemistries may not be represented.

There are more specific considerations as well which impact generalizability. Datasets can be noisy with large error bars or biased toward certain structures or endpoint values [40]. Noisy datasets are an issue because noise decreases the reliability of each of the datapoints, which decreases the reliability of the model as a whole. The model cannot be more accurate than the experimental data it was trained on. So, with any assay data source, the robustness and the reliability of the assay must be carefully assessed in tandem with domain experts to determine how it can be used in model-building.

Systematic biases in assay design also effect generalizability. For example, in fluorescent assays, compounds can interfere with the readouts by either emitting (autofluorescence) or absorbing the appropriate wavelength of light (quenching), leading to systematic biases in the data and thus a systematically biased model, where auto-fluorescent and quenching compounds would not be in its applicability domain [78]. Furthermore, the inaccurate data points decrease the reliability of the model, because the model may find non-biologically relevant patterns from the inaccurate data. There are strategies to mitigate these effects, including using another distinct AI model to find and isolate the inaccurate data points, but the broader point is that these concerns are complex and require a nuanced understanding of the underlying system.

Drug discovery data also has inherent biases related to the iterative design process. There is a compound series bias [79], where for any given endpoint the data available will all be built off of the same few series, simply because those happened to be the ideas a team was exploring. Thus, for novel compounds, which are the ones we are interested in, the model's prediction may be unreliable. There is also a cascade bias, where only compounds which make it past the first screen of assays are assessed in the second screen. Thus, a model which is trained only on data from the second screen may be unreliable when it comes to compounds that would not make it past the first screen.

Another key issue is a range bias, where the desired output property is often out of the range of the experimental data. For example, suppose we are training a model on inhibition against a certain target without structure data and we are looking for highly potent compounds, but we have not found any such compounds yet. The AI model's prediction would not be reliable because it has not seen any examples of highly potent compounds it could use to make a prediction.

3.3. Dataset construction

In recent years, there has been a large increase in the volume of drug discovery-related data both in industry and the public domain [80]. Published data sets are stored in databases such as PubChem [74], ChEMBL [81], and ZINC [82] and are available for general use. Pharmaceutical companies have their own internal proprietary databases which are rich in compound diversity and data sets. There has been an important effort to pool these proprietary databases together, retaining the information value while obscuring the underlying proprietary compounds, in order to leverage the sum of knowledge to create larger, more comprehensive databases to build more accurate models [83–86]. The largest of these initiatives is MELLODDY, a consortium including 10 pharmaceutical partners, which uses federated learning (FL) to train an AI model using decentralized proprietary data.

However, more often than not this data is heterogenous, such that a model cannot be straightforwardly applied [30,87]. Instead, data must be carefully curated to create a dataset for model development. In the curation process, great care must be taken to ensure that data come from the same assay with the same parameters and if aggregating data, domain experts must be engaged to ensure that the assays are functionally the same. Furthermore, the data normalization procedures must be applied uniformly to the entire dataset. If the available data consists of multiple assays measuring the same fundamental information in different ways (i.e. scale or units), the data may be used to create a classification model instead of regression model.

After data curation is performed, the selection of data splits – creating train, test, and validation datasets – is also crucial toward producing generalizable models [88]. If the dataset is split improperly, the quality of the model cannot be appropriately judged using the test set, such that the model cannot be deployed with high confidence [89]. This impacts the decidability of a model. Sheridan [79] recommends creating data splits using the timestamps of when each data point is collected, finding that using a random-selection for data was overly optimistic in predicting the practical accuracy of AI models, while using a leave-class-out selection was overly pessimistic. In a practical evaluation and deployment of AI model, data splits are crucial as a misperception of the accuracy model can lead to a team either being misled down a wrong direction or underutilizing the value of the model.

4. Learning frameworks

To overcome many of these data issues, transfer learning (Section 4.1), multitask learning (Section 4.2), and meta learning (Section 4.3) frameworks have been explored. Figure 2 presents a graphical description of the aforementioned frameworks which are further defined in the body of the article in their respective sections. Each of these frameworks is built on and requires the veracity of the hypothesis that molecular property prediction tasks are all similar to each other, so by giving an AI model extra information about other tasks can make it better at some given tasks. The extra information makes the model quite reliable. One fundamental reason this is true is that all molecular property prediction tasks are rooted in some physical, chemical, biological system, so if an AI model which uses information from other tasks to learn physics, biology and chemistry, it will do better on any given task.

4.1. Transfer learning

With transfer learning, a model generalizes knowledge from one task to another to increase its applicability and decidability. This approach has already been used to an extent in the practice of tweaking global endpoint models to create local models, and the recent trends have been in working toward incorporating data beyond that. The two most common transfer learning approaches are feature-based, where one model learns some molecular representation which is then used in other models, and parameter-based, where one model is trained on its task and then its weights are used as an approximate solution and fine-tuned to create a model on a different task [90–92].

Using a feature-based transfer learning, Abbasi et al. [93] demonstrated that appropriately transferring features learned from one dataset to another dataset improves the model's performance. They found that the more related the tasks were, the greater the benefit. For example, knowledge of ToxCast data improved Tox21 model performance more than it did SIDER performance, and while both Tox21 and ToxCast describe the toxicity of compounds, SIDER describes drug disorders. ToxCast expands the applicability and reliability of predictions more on Tox21 then on SIDER because the information it gives is more similar and more directly useful. Wang et al. [94] also used featured-based transfer learning, with a Transformer model giving the learned feature

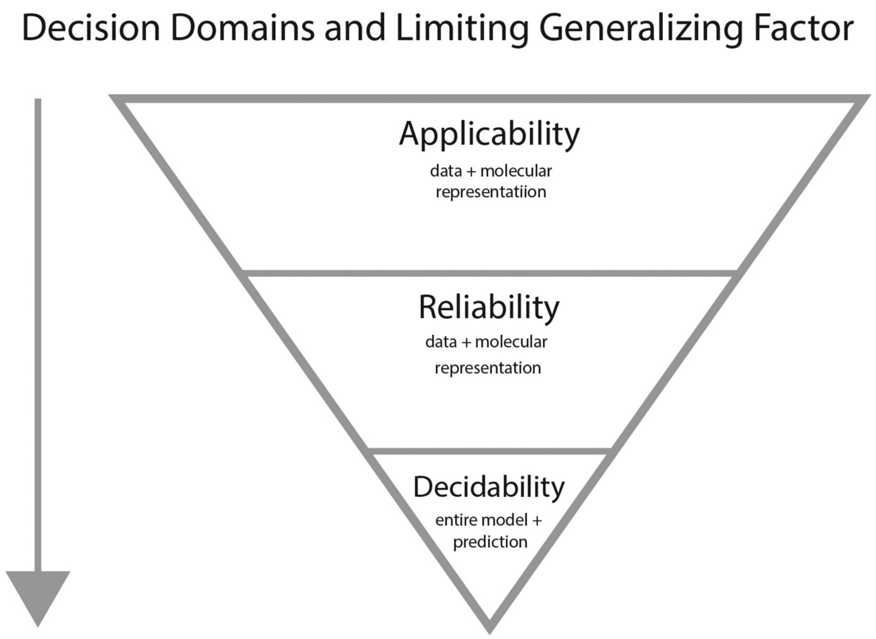
Decision Domains and Limiting Generalizing Factor
Applicability
data + molecular representatiion
Reliability
data + molecular representation
Decidability
entire model + prediction

Figure 1. The three-staged decision domain hierarchy and their limiting generalizing factor.

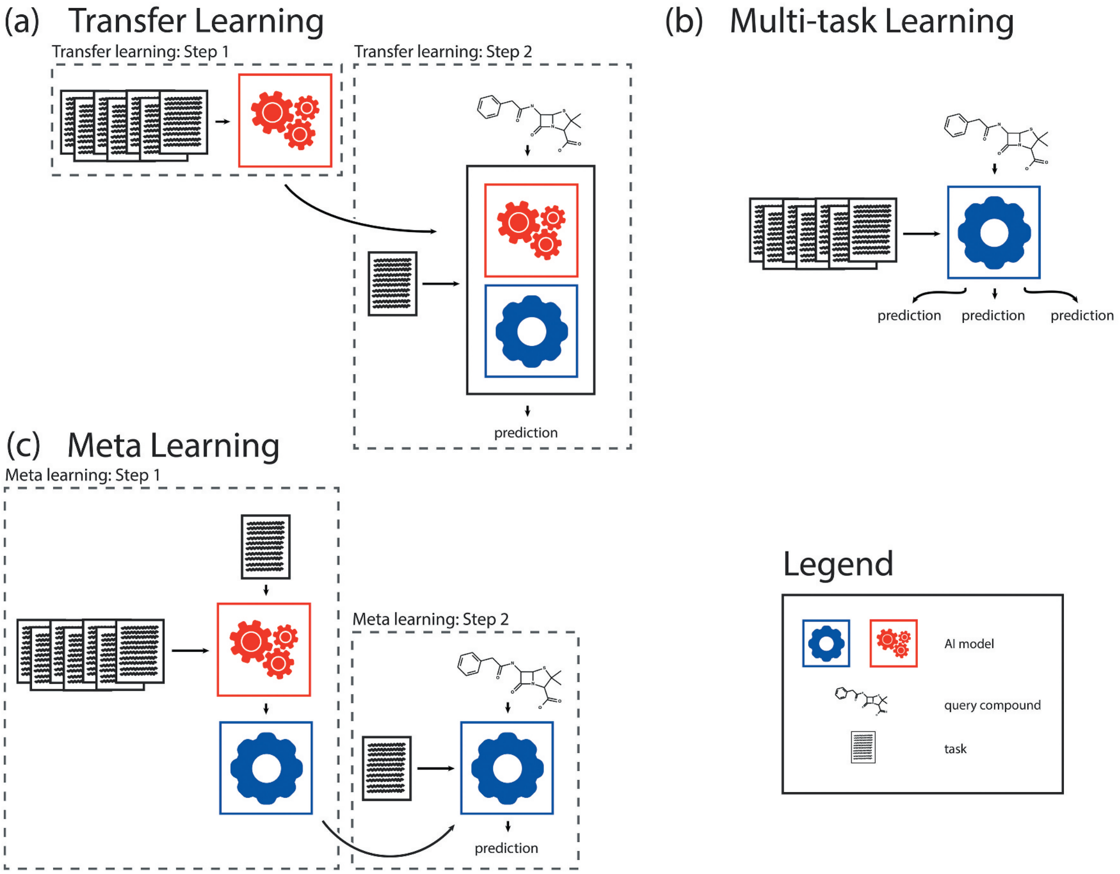

Figure 2. The figure presents graphical depiction of the following: (a) transfer learning models – described in Section 4.1 – where in step 1, a general AI model is trained on a large collection of data, and where in step 2, the general AI model is refined for use on a specific task; (b) multitask learning models – described in Section 4.2 – where an AI model is trained to predict multiple different properties or endpoints at the same time; and (c) meta learning models – described in Section 4.3 – where a second-order AI model is trained such that for a given task it outputs model specifications for that task, and then a model with those specifications is trained on the task.

representation, and achieved better results than using circular fingerprints on predicting protein inhibition.

Li et al. found that by using a parameter-based transfer learning approach, they could improve the performance of a baseline model on predicting Log P, solvation energy, HIV replication inhibition, and blood-brain barrier penetration [95]. Their model was an adaptation of the ULMFiT, a natural language processing algorithm, indicating the universality of transfer learning approaches.

4.2. Multitask learning

Multitask learning is where a single AI model predicts for one compound many different properties at the same time [96]. The core concept behind this framework is that the AI model will share information across each of the properties it is predicting. In terms of applicability domains, suppose we have some compound which we want to predict some property for, and that compound has no similar compounds with experimental data. If we were training a single-task model it would be out of the AD. But with multitask learning, the hypothesis is that there may be experimental data for similar compounds on different properties, whose information the AI model can use to make an informed prediction.

In 2018, Mayr et al. showed that using multitask learning with an MLP on molecular descriptors to predict the drug target activity of 1,310 assays from ChEMBL resulted in a model that not only outperformed all other methods (including SVMs and RFs) but also in limited circumstances reached near experimental accuracy [21]. In a recent 2020 study, Feinberg et al. [97,98] found that multitask learning with an MPNN on a molecular performance worked significantly better on a diverse set of ADME-T prediction tasks versus both a single-task variant as well as an RF baseline model, and similar simplified versions have been successfully used in industry [99].

4.3. Meta learning

In the meta-learning framework, a second-order model is created where the second-order model takes in the data for the task and then outputs the specifications for another AI model which is then trained on the task [100]. The specifications may be the type of model, hyperparameters to use, or weight initializations. This technique helps because certain models do better on certain tasks – as we have seen already on larger datasets, deep learning models work better, but on smaller datasets traditional techniques such as SVMs work better.

In a 2018 study, Olier et al. showed that their Meta-QSAR model outperformed a random forest model by 13% on a subset of ChEMBL for drug target activity prediction [101]. In another approach, Nguyen et al. used a meta-learning model to create weight-initializations for a given task which were then trained on. They validated their approach on the same ChEMBL dataset as Mayr et al. with the large caveat that they used only tasks with at least 128 data points [21] and found that in that case, the meta-learning model outperformed multitask models [102].

5. Conclusion

AI generated in silico predictive models for key ADME/Tox endpoints and activities are of high value to the early drug discovery process as they will expedite the selection of safer and more efficacious drugs for the clinic, which will ultimately improve patient lives at a reduced economic cost. One key challenge that remains is the availability of data and the generalizability of the models that depend on the data.

Data must be carefully evaluated for bias or noise and then uniformly and carefully processed in order to construct robust datasets. After constructing the datasets, models must be developed to best generalize from the data and make appropriate predictions. Each model consists of a molecular representation input into an AI algorithm trained on a learning framework. Although currently the choice of which components to use to create the model has no clear answer, the general trend is toward a learned features representation and a deep learning model, which are integrated under some meta learning, transfer learning, or multitask learning framework.

6. Expert opinion

The field of AI is making an impact across many industries. There is no doubt that AI will also have an impact in the discovery of new medicines. One main area where AI predictive models can be particularly effective is in the early drug discovery phase where clinical candidates are nominated. The ability to apply AI to expedite and reduce the cost in the discovery of new medicines to help improve patient lives is priceless. To this end, many companies are working on leveraging proprietary data as well as publicly available datasets generated for ADME/Tox endpoints retrospectively. We propose that there should also be more of a commitment in both academics and industry to the prospective generation data. This may be viewed as challenging and costly; after all, it may be unrealistic to standardize primary endpoint assays within a global company let alone in different companies and it would be resource prohibiting to run primary endpoint assays on every compound that is registered in a company database. However, this commitment to collecting data is an essential component to develop robust and reliable AI-driven predictive tools. If both academic institutions and drug discovery companies commit to sharing data in a thoughtful manner, then there will be significant value added for generating predictive models by combining data and removing siloes. But the proprietary nature of compounds is always a concern in industry. To deidentify shared compounds, shared datasets can be reduced to a database of features and related descriptors. However, in reality, collecting and sharing data is still an aspirational goal for drug discovery in both academia and industry and will be challenging to achieve. Furthermore, even with large datasets, we still want AI models to be as efficient as possible in extracting information and making predictions.

The ultimate solution lies in the ability of AI models to become more generalizable. This becomes an even more important consideration as the drug discovery space is moving beyond small molecules to new modalities as therapeutics, where AI must in fact extrapolate to the broader drug space. The ability of AI to generalize has been proven in domains such as computer vision, natural language processing, and games. And with the recent success of AlphaFold2 in protein structure prediction from sequences, it is clear that AI is able to understand and make accurate calculations at the biochemical level. Thus, as ADME/Tox endpoint and activity prediction study similar fundamental systems, we can expect that AI models will be similarly successful. Indeed, many recent studies have demonstrated the success of models in predicting diverse endpoints such as blood-brain barrier penetration and drug-target activity. However, although the potential benefits are clearly evident, the rapid pace of development and volume of studies means that the practical choice of model and dataset designed is often unclear for the practical industry practitioner. What is clear, however, is that AI models will become larger with more learnable parameters and learned feature representation, and the AI model of the future will be able to learn basic biological, chemical, and physical principles from large datasets and apply that knowledge to accurately predict some ADME/Tox endpoint or activity, even if there is low experimental data availability for that specific task. However, given the historical evidence of AI in other domains, the best AI model 5 years from now has likely not been discovered yet. To get there however, researchers must undertake systematic studies with a careful eye toward the nuances of drug discovery data.

That future is bright. Ideally, just as project teams use experimental data generated from *in vitro* and *in vivo* assays to assess the quality of drug candidates, *in silico* model predictions will also be incorporated as first-pass triage of compounds. Given the rise in computational resources, AI models can be used almost as ultra-efficient high-throughput screens, allowing teams to both increase the quantity of compounds they consider, and the quality of compounds going to experiment. However, experimental data will always be critical to providing the ground-truth data and guiding the development of such models.

Acknowledgments

The authors thank Dr S Szalma (Takeda Global Head of Computational Biology), Dr. L Hamann (Takeda Head of Drug Discovery Sciences), and Professor I Tsigelny (San Diego Supercomputer Center) for their critical reading of the manuscript.

Funding

This manuscript was not funded.

Declaration of interest

SS Bahmanyar and JC Baber are both employees of Takeda Pharmaceuticals. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
2. Brown TB, Mann B, Ryder N; et al. Language models are few-shot learners. *ArXiv200514165 Cs*, 2020.
3. Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks and tree search. Nature. 2016;529(7587):484–489.
4. Goodfellow I, Bengio Y, Courville A. *Deep learning*; adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press; 2016.

- **A comprehensive book on the fundamentals of deep learning.**

5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

- **A review encompassing main themes and advancements in deep learning.**

6. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477.

- **A review on the key areas where AI will transform drug discovery.**

7. Schneider P, Walters WP, Plowright AT, et al. Rethinking drug design in the artificial intelligence Era. Nat Rev Drug Discov. 2020;19(5):353–364.

- **A perspective presenting a realistic discussion on the coming grand challenges in AI drug discovery.**

8. Freedman DH. Hunting for new drugs with AI. Nature. 2019;576(7787):S49–S53.
9. Yang X, Wang Y, Byrne R, et al. Concepts of artificial intelligence for computer-assisted drug discovery. Chem Rev. 2019;119(18):10520–10594.
10. Properzi F, Taylor K, Steedman M, et al. Intelligent drug discovery; Deloitte Centre for Health Solutions; Deloitte University, B-1831 Diegem, Berkenlaan. 2019.

•• **A report with a business focused analysis on AI's role in guiding drug discovery.**

11. *A*I for drug discovery, biomarker development, and advanced R&D landscape overview 2019/Q2; Deep Knowledge Analytics "Pharma Division". 2019. [cited 2021 Dec 31]. Available from: https://ai-pharma.dka.global/quarter-2-2019/.

•• **An overview of the major industry players in drug discovery and AI.**

12. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702.e13.
13. Callaway E. 'It will change everything': deepMind's AI makes gigantic leap in solving protein structures. Nature. 2020;588(7837):203–204.

•• **A news article on DeepMind's AlphaFold2's groundbreaking success in protein structure prediction.**

14. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577(7792):706–710.
15. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435–441.
16. Unterthiner T, Mayr A. Deep learning as an opportunity in virtual screening. In Advances in Neural Information Processing Systems; Curran Associates, Inc.: Redhook NY USA. 2014; 27.

17. Struble TJ, Alvarez JC, Brown SP, et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. J Med Chem. 2020;63(16):8667–8682.
18. Walters WP, Barzilay R. Applications of deep learning in molecule generation and molecular property prediction. Acc Chem Res. 2020; acs.accounts.0c00699. DOI: 10.1021/acs.accounts.0c00699.
- **A recent review on DL techniques applied to *de novo* design and property prediction.**
19. Merck molecular activity challenge. cited [2020 Nov 29]. Available from: https://kaggle.com/c/MerckActivity
20. Ma J, Sheridan RP, Liaw A, et al. Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inf Model. 2015;55(2):263–274.
21. Mayr A, Klambauer G, Unterthiner T, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci. 2018;9(24):5441–5451.
22. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(D1): D1100–D1107.
23. Jordan AM. Artificial intelligence in drug design—the storm before the calm? ACS Med Chem Lett. 2018;9(12):1150–1152.
- **A tempered analysis of the practical trajectory of AI and the challenges in its application towards drug discovery.**
24. Xu Y, Li X, Yao H, et al. Neural networks in drug discovery: current insights from medicinal chemists. Future Med Chem. 2019;11 (14):1669–1672.
25. Su M, Feng G, Liu Z, et al. Tapping on the black box: how is the scoring power of a machine-learning scoring function dependent on the training set? J Chem Inf Model. 2020;60 (3):1122–1136.
- **An analysis on the impact of dataset composition on the reliability and predictive power of machine learning methods.**
26. Liu R, Wang H, Glover KP, et al. Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks? J Chem Inf Model. 2019;59 (1):117–126.
- **An analysis focused on determining the applicability domain of deep neural networks.**
27. Paul D, Sanap G, Shenoy S, et al. Artificial intelligence in drug discovery and development. Drug Discov Today. 2020;26(1): S1359644620304256.
28. Brown N, Ertl P, Lewis R, et al. Artificial intelligence in chemistry and drug design. J Comput Aided Mol Des. 2020;34(7):709–715.
29. Brown N, Cambruzzi J, Cox PJ, et al. Big data in drug discovery. In: Progress in Medicinal Chemistry. Vol. 57, Elsevier; 2018. p. 277–356.
30. Zhu H. Big data and artificial intelligence modeling for drug discovery. Annu Rev Pharmacol Toxicol. 2020;60(1):573–589.
- **A review encompassing data quality and dataset construction, emphasizing the use of public datasets and datasharing.**
31. Irwin BWJ, Levell JR, Whitehead TM, et al. Practical applications of deep learning to impute heterogeneous drug discovery data. J Chem Inf Model. 2020;60(6):2848–2857.
- •• **A research article on the effectiveness of deep learning for data imputation, filling out missing data fields.**
32. Zhang Y, Yang QA Survey on multi-task learning. *ArXiv170708114 Cs*, 2018.
33. Pan SJ, Yang QA. Survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–1359.
- •• **An AI-focused survey on transfer learning and it's fundamental principles.**
34. Brigato L, Iocchi L, Close A Look at deep learning with small data. *ArXiv200312843 Cs Stat*, 2020.
35. David L, Thakkar A, Mercado R, et al. Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminformatics. 2020;12(1):56.
- **A review on molecular representations, giving practical considerations on where to use what representation.**
36. Lo YC, Rensi SE, Torng W, et al. Machine Learning in Chemoinformatics and Drug Discovery. Drug Discovery Today, 2018;23(8):1538–1546. Available from: https://doi.org/10.1016/j.drudis.2018.05.010
37. Durant JL, Leland BA, Henry DR, et al. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci. 2002;42 (6):1273–1280.
38. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754.
39. Fieser LF, Ettlinger MG, Fawaz G. Naphthoquinone antimalarials. XV. Distribution between organic solvents and aqueous buffers 1,2. J Am Chem Soc. 1948;70(10):3228–3232.
- **A historical summary and future outlook on QSAR modelling, discussing the nuances of working with drug discovery data.**
40. Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57 (12):4977–5010.
41. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics: volume i: alphabetical listing/volume II: appendices, references. 1st ed. Methods and Principles in Medicinal Chemistry. Vol. 41, Wiley; 2009. DOI: 10.1002/9783527628766
42. Topliss JG, Costello RJ. Chance correlations in structure-activity studies using multiple regression analysis. J Med Chem. 1972;15 (10):1066–1068.
- **A key, early study analyzing the generalizability of machine learning approaches with respect to their molecular representations.**
43. Koutsoukas A, Paricharak S, Galloway WRJD, et al. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. J Chem Inf Model. 2014;54 (1):230–242.
44. Bender A, Jenkins JL, Scheiber J, et al. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model. 2009;49(1):108–119.
45. Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res. 2009;20(3–4):241–266.
46. Chuang KV, Gunsalus LM, Keiser MJ. Learning molecular representations for medicinal chemistry: miniperspective. J Med Chem. 2020;63(16):8705–8722.
- **A recent review on the use of deep learning for learned features.**
47. Ponzoni I, Sebastián-Pérez V, Requena-Triguero C, et al. Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery. Sci Rep. 2017;7(1):2403.
48. Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370–3388.
- •• **A recent study highlighting and analyzing the performance of their Directed Message Passing Neural Network on diverse properties and endpoints.**
49. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513–530.
- •• **A comprehensive comparison of many AI methodologies for property prediction and a useful framework for the development of AI models.**
50. Korotcov A, Tkachenko V, Russo DP, et al. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol Pharm. 2017;14 (12):4462–4475.
51. Koutsoukas A, Monaghan KJ, Li X, et al. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminformatics. 2017;9(1):42.
52. Tsou LK, Yeh S-H, Ueng S-H, et al. Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR Agonist discovery. Sci Rep. 2020;10(1):16771.
53. Sejnowski TJ. The unreasonable effectiveness of deep learning in artificial intelligence. Proceedings of the National Academy of Sciences. 2020;117(48):30033–30038.

54. Gilmer J, Schoenholz SS, Riley PF, et al. Neural Message Passing for Quantum Chemistry. In Proceedings of the 34th International Conference on Machine Learning; Proceedings of Machine Learning Research; PMLR: International Convention Centre, Sydney, Australia, 2017;70:1263–1272.

55. Li Y, Tarlow D, Brockschmidt M, et al. Gated graph sequence neural networks. *ArXiv151105493 Cs Stat*, 2017.

56. Schütt KT, Arbabzadah F, Chmiela S, et al. Quantum-chemical insights from deep tensor neural networks. Nat Commun. 2017;8 (1):13890.

57. Chakravarti SK, Alla SRM. Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. Front Artif Intell. 2019;2:17.

58. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model. 2013;53(7):1563–1575.

59. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *ArXiv170603762 Cs*, 2017.

60. Öztürk H, Özgür A, Schwaller P, et al. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discov Today. 2020;25(4):689–705.

61. Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci. 2019;5(9):1572–1583.

62. Nayak P, Silberfarb A, Chen R, et al. Transformer based molecule encoding for property prediction. *ArXiv201103518 Q-bio*, 2020.

63. Netzeva TI, Worth AP, Aldenberg T, et al. Current status of methods for defining the applicability domain of (Quantitative) structure-activity relationships: the report and recommendations of ECVAM workshop 52. Altern Lab Anim. 2005;33(2):155–173.

64. Maple HJ, Clayden N, Baron A, et al. Developing degraders: principles and perspectives on design and chemical space. MedChemComm. 2019;10(10):1755–1764.

65. Chopra R, Sadok A, Collins I, et al. Evaluation of the approaches to targeted protein degradation for drug discovery. Drug Discov Today Technol. 2019;31:5–13.

66. Costales MG, Childs-Disney JL, Haniff HS, et al. How we think about targeting RNA with small molecules. J Med Chem. 2020;63 (17):8880–8900.

67. Hanser T, Barber C, Guesné S, et al. Applicability domain: towards a more formal framework to express the applicability of a model and the confidence in individual predictions. Advances in Computational Toxicology. Hong H, editor. Vol. 30, Cham: Challenges and Advances in Computational Chemistry and Physics; Springer International Publishing; 2019. 215–232.

- **A formal treatment and analysis of applicability domains of QSAR models.**

68. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptor space: a review. Altern Lab Anim. 2005;33 (5):445–459.

69. Mathea M, Klingspohn W, Baumann K. Chemoinformatic classification methods and their applicability domain. Mol Inform. 2016;35 (5):160–180.

•• **A comprehensive review of natural applicability domain metrics for traditional machine learning methods.**

70. Carrió P, Pinto M, Ecker G, et al. Applicability Domain Analysis (ADAN): a robust method for assessing the reliability of drug property predictions. J Chem Inf Model. 2014;54(5):1500–1511.

•• **A analysis of general applicability domain metrics for evaluating the reliability of models.**

71. Klingspohn W, Mathea M, Ter Laak A, et al. Efficiency of different measures for defining the applicability domain of classification models. J Cheminformatics. 2017;9(1):44.

72. Hirschfeld L, Swanson K, Yang K, et al. Uncertainty quantification using neural networks for molecular property prediction. J Chem Inf Model. 2020;60(8):3770–3780.

73. Eklund M, Norinder U, Boyer S, et al. The application of conformal prediction to the drug discovery process. Ann Math Artif Intell. 2015;74(1):117–132.

74. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–D1109.

75. Wong SC, Gatt A, Stamatescu V, et al. Understanding data augmentation for classification: when to warp? In 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA); IEEE: Gold Coast, Australia, 2016; pp 1–6. DOI: 10.1109/DICTA.2016.7797091.

76. Bjerrum EJ SMILES enumeration as data augmentation for neural network modeling of molecules. arXiv:1703.07076 [cs], 2017.

77. Asilar E, Hemmerich J, Ecker GF. Image based liver toxicity prediction. J Chem Inf Model. 2020;60(3):1111–1121.

78. Borrel A, Huang R, Sakamuru S, et al. High-throughput screening to predict chemical-assay interference. Sci Rep. 2020;10(1):3986.

79. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model. 2013;53 (4):783–790.

80. Rifaioglu AS, Atas H, Martin MJ, et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform. 2019;20 (5):1878–1912.

81. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945–D954.

82. Irwin JJ, Tang KG, Young J, et al. ZINC20—A free ultralarge-scale chemical database for ligand discovery. J Chem Inf Model. 2020;60 (12):6065–6073.

83. Pejó B The good, the bad, and the ugly: quality inference in federated learning. *ArXiv200706236 Cs Stat*, 2020.

84. Simpson PB, Wilkinson GF. What makes a drug discovery consortium successful? Nat Rev Drug Discov. 2020;19(11):737–738.

85. Hinkson IV, Madej B, Stahlberg EA. Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. Front Pharmacol. 2020;11. DOI:10.3389/fphar.2020.00770

•• **An article describing the importance of successful collaboration and teamwork in the application of machine learning towards drug discovery.**

86. Year 1 announcement. [cited 2021 Mar 1]. Available from: https://www.melloddy.eu/y1announcement

87. Wild DJ. Mining large heterogeneous data sets in drug discovery. Expert Opin Drug Discov. 2009;4(10):995–1004.

88. Riley P. Three Pitfalls to avoid in machine learning. Nature. 2019;572(7767):27–29.

89. D'Amour A, Heller K, Moldovan D; et al. Underspecification presents challenges for credibility in modern machine learning. *ArXiv20 1103395 Cs Stat*, 2020.

90. Cai C, Wang S, Xu Y, et al. Transfer learning for drug discovery. J Med Chem. 2020;63(16):8683–8694.

- **A comprehensive, recent review on the use of transfer learning in drug discovery.**

91. Duvenaud DK, Maclaurin D, Iparraguirre J, et al. InAdvances in Neural Information Processing Systems; Curran Associates, Inc.: Redhook NY USA. 2015;28.

92. Coley CW, Barzilay R, Green WH, et al. Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model. 2017;57(8):1757–1772.

93. Abbasi K, Poso A, Ghasemi J, et al. Deep transferable compound representation across domains and tasks for low data drug discovery. J Chem Inf Model. 2019;59(11):4528–4539.

•• **A recent study on the use of an integrated, deep learning for transfer learning for low data ADME/Tox endpoint prediction.**

94. Wang S, Guo Y, Wang Y, et al. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*; ACM: Niagara Falls NY USA, 2019; pp 429–436. DOI: 10.1145/3307339.3342186.

- •• **A recent study on the use of AI transformers for unsupervised transfer learning.**

95. Li X, Fourches D. Inductive transfer learning for molecular activity prediction: next-Gen QSAR models with MolPMoFiT. J Chemin formatics. 2020;12(1):27.
96. Ramsundar B, Liu B, Wu Z, et al. Is multitask deep learning practical for pharma? J Chem Inf Model. 2017;57(8):2068–2076.

- • **An outlook on the relevance and importance of multitask learning.**

97. Feinberg EN, Joshi E, Pande VS, et al. Improvement in ADMET prediction with multitask deep featurization. J Med Chem. 2020;63(16):8835–8848.

- •• **A recent study demonstrating the effectiveness of multitask and graph-based deep learning for ADME/Tox endpoint prediction.**

98. Feinberg EN, Sur D, Wu Z, et al. PotentialNet for molecular property prediction. ACS Cent Sci. 2018;4(11):1520–1530.
99. Montanari F, Kuhnke L A multitask deep learning model for physico-chemical property prediction. *Gordon Research Conference in Computer Aided Drug Design*. 2019.
100. Finn C, Abbeel P, Levine S Model-Agnostic meta-learning for fast adaptation of deep networks. *ArXiv170303400 Cs*, 2017.

- • **A framework for the use of meta-learning.**

101. Olier I, Sadawi N, Bickerton GR, et al. Meta-QSAR: a large-scale application of meta-learning to drug design and discovery. Mach Learn. 2018;107(1):285–311.

- •• **An article describing the successful application of AI toward ADME/Tox endpoint and property prediction.**

102. Nguyen CQ, Kreatsoulas C, Branson KM Meta-learning GNN initializations for low-resource molecular property prediction. *ArXiv200305996 Phys. Stat*, 2020.