

## Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery

Xin Yang,<sup>†</sup> Yifei Wang,<sup>†</sup> Ryan Byrne,<sup>‡</sup> Gisbert Schneider,<sup>\*,‡</sup> and Shengyong Yang<sup>\*,†</sup>

<sup>†</sup>State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

<sup>‡</sup>ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, CH-8093 Zurich, Switzerland

**ABSTRACT:** Artificial intelligence (AI), and, in particular, deep learning as a subcategory of AI, provides opportunities for the discovery and development of innovative drugs. Various machine learning approaches have recently (re)emerged, some of which may be considered instances of domain-specific AI which have been successfully employed for drug discovery and design. This review provides a comprehensive portrayal of these machine learning techniques and of their applications in medicinal chemistry. After introducing the basic principles, alongside some application notes, of the various machine learning algorithms, the current state-of-the-art of AI-assisted pharmaceutical discovery is discussed, including applications in structure- and ligand-based virtual screening, de novo drug design, physicochemical and pharmacokinetic property prediction, drug repurposing, and related aspects. Finally, several challenges and limitations of the current methods are summarized, with a view to potential future directions for AI-assisted drug discovery and design.



### CONTENTS

1. Introduction
  - 1.1. Brief History of Artificial Intelligence
  - 1.2. Brief Introduction to the History of Artificial Intelligence Application in Drug Discovery
  - 1.3. Goal, Scope, And Organization of This Review
2. How Artificial Intelligence Works
  - 2.1. Basic Principles
    - 2.1.1. Classes of Learning Tasks and Techniques
    - 2.1.2. Learning Algorithms
  - 2.2. Artificial Intelligence Modeling in Drug Discovery
    - 2.2.1. Constructing a Model
    - 2.2.2. Input Data Preparation
3. Applications of Artificial Intelligence in Structure-Based Virtual Screening
  - 3.1. Constructing Artificial Intelligence-Based Nonpredetermined Scoring Functions
  - 3.2. Structure-Based Virtual Screening with Artificial Intelligence Approaches
4. Artificial intelligence approaches in ligand-based virtual screening
  - 4.1. Artificial Intelligence-Based QSAR Models
  - 4.2. Other Artificial Intelligence-Assisted Classification Strategies
5. Applications of Artificial Intelligence in De Novo Drug Design
6. Physicochemical and ADMET Properties Predictions
  - 6.1. Prediction of Physicochemical Properties

B	6.1.1. Lipophilicity Prediction	AJ
B	6.1.2. Aqueous Solubility Prediction	AK
D	6.1.3. Intrinsic Permeability Prediction	AL
D	6.1.4. Predictions of Other Physicochemical Properties	AM
D	6.2. ADMET Predictions	AM
F	6.2.1. Absorption	AM
F	6.2.2. Distribution	AM
F	6.2.3. Metabolism	AN
F	6.2.4. Excretion	AP
F	6.2.5. Toxicity	AP
H	7. Applications of Artificial Intelligence in Drug Repurposing	AS
Q	7.1. Drug Repurposing under the Supervised Learning Paradigm	AS
Q	7.2. Drug Repurposing under the Unsupervised Learning Paradigm	AT
S	7.3. Drug Repurposing under the Semisupervised Learning Paradigm	AU
T	8. Applications of Artificial Intelligence in Other Aspects of Drug Discovery	AV
U	8.1. Precision Medicine and Big Data-Driven Drug Discovery	AV
W	8.2. Artificial Intelligence-Assisted Molecular Simulations	AW
W	8.3. Data-Driven Explorations of Polypharmacology	AX
Y	8.4. Artificial Intelligence-Assisted Evaluation of Synthetic Accessibility	AY
AA		
AI		
AJ		

Received: November 29, 2018

9. Concluding Remarks and Future Perspectives	AZ
Author Information	BA
Corresponding Authors	BA
ORCID	BA
Notes	BA
Biographies	BA
Acknowledgments	BB
Abbreviations	BB
References	BB

## 1. INTRODUCTION

The research and development (R&D) cycle for innovative small-molecule drugs faces many challenges, such as high cost-to-market, limited success in clinical trials, and long cycle times.<sup>1,2</sup> The productivity of drug R&D (herein referring exclusively to small molecule drugs, unless otherwise specified) in the pharmaceutical industry remains on the decline, despite record expenditures.<sup>3,4</sup> The reasons for this tendency are manifold, including current market saturation, difficulties in bringing novel chemical matter through a complex approvals process, and the willingness-to-pay in developed and developing markets, among others. Here, we will focus on a central challenge in medicinal chemistry, namely the inherent difficulty of translating the process of drug discovery from basic science to early clinical trials. Today, scientists have more information than ever before on a range of topics pertinent to the subject matter, far outpacing the ability of most to properly parse and integrate into their own workflows and research objectives. One solution to such problems is to “outsource” our reasoning to a machine intelligence when it comes to the analysis of multisource and multidimensional data. In this context, machine learning and domain-specific (“weak”) artificial intelligence (AI) offer fresh opportunities for small-molecule drug discovery. Machine learning approaches that might be considered instances of weak AI have made remarkable progress, with developments in both their fundamental algorithms and applications. Thus, we will use the term “AI” as a synonym for certain machine learning techniques because there is no “strong” (general) AI to date. Therefore, this review will focus on those aspects of this promising subfield that have already demonstrated their usefulness and applicability and reflect on those technologies which seem most promising for the next phase of AI in drug discovery.

### 1.1. Brief History of Artificial Intelligence

The term “artificial intelligence” was coined by John McCarthy at the Dartmouth Conference in 1956 to describe “the science and engineering of making intelligent machines”.<sup>5</sup> McCarthy’s original description still holds true today, albeit with some fleshing-out of the specifics. As a multidisciplinary field, AI involves integrating insights from diverse disciplines such as computer science, mathematics, psychology, linguistics, philosophy, neuro-science, artificial psychology, and many others. Recent intellectual and engineering advances have helped the field progress from purely theoretical studies to the implementation of intelligent systems that solve problems in various aspects of our lives. The current scope of such applications includes fields and studies as heterogeneous as natural language understanding and processing, speech understanding and processing, mechanical/computer vision, autonomous/intelligent robots, and domain expertise acquisition, to

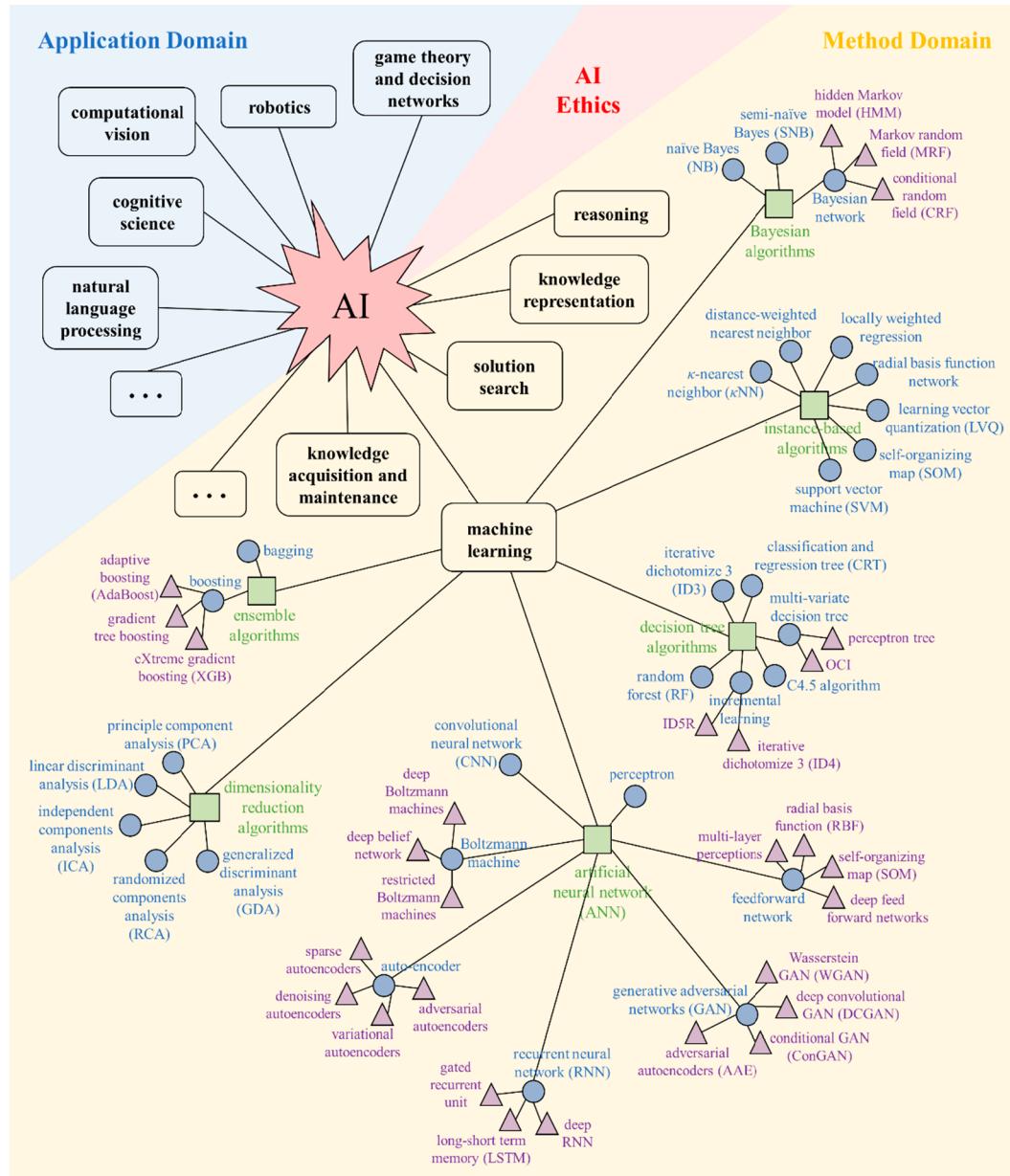
provide only a few examples (Figure 1). Despite the broad spectrum of problems that can be addressed by AI, there are some basic methods that play major roles in all cases,<sup>6,7</sup> examples of which include knowledge acquisition and maintenance, knowledge representation, solution search, logic reasoning, and machine learning.

In the mid-1930s, Alan Turing introduced the idea of what is today referred to as the “universal Turing machine”, which could simulate any possible computer. To some extent in march-step with the advances of the computer hardware, the history of AI has been one of fits and starts, of boom and bust (Figure 2). The early heyday, in the 1950s and 1960s, was fueled by pure optimism. During this period, symbolic methods were introduced for semantic processing, the concepts of logical reasoning and heuristic searching emerged, and man-machine interaction became feasible. The first machines with preliminary intelligence were conceived, for example, STUDENT (1964),<sup>8</sup> a machine that could implement machine proofs of some mathematical theorems and logical inference of statements. Another early example was ELIZA (1966),<sup>9</sup> a machine that could emulate human dialogue, albeit in a limited fashion. The rapid development of these and other instances of AI fueled a frothy reaction, leading to a cycle of irrational exuberance and eventual disappointment in the power of AI. The eventual cooling of such sentiments has been described as the first “AI winter”. Importantly, and of note for AI evangelists past and present, there is no magic in AI, only probability and statistics, the proper applications of which are contingent on mathematics, the availability of suitable data, and on the capabilities of our hardware.

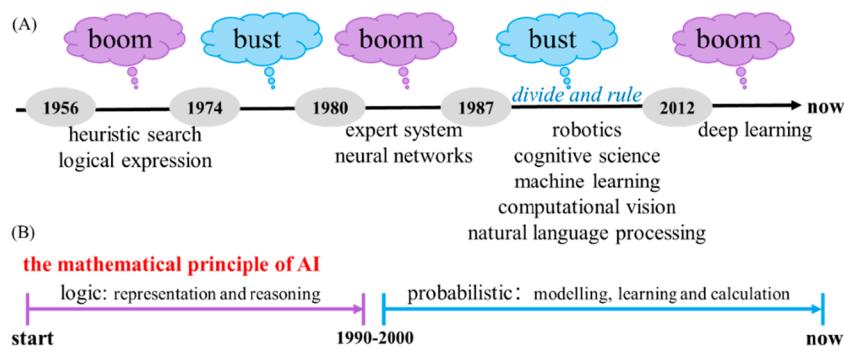
AI had its second peak in the early 1980s. Substantial progress had been made in AI-related mathematical models, including the multilayer feed-forward neural network<sup>10,11</sup> and the backpropagation algorithm.<sup>12</sup> These tools allow for the construction of an abstract model of the world and provide a way to update the model given input (learning from feedback). This combination was the first to manage victory against a human chess player, and one that laid the foundation for much of the work done in the field to date.

A first foray of such methods into the area of chemistry and molecular biology was achieved by the prediction of secondary structure from protein sequence information.<sup>13</sup> At the same time, various expert systems had entered the market. For example, Carnegie Mellon University had created an expert system for the DEC Company.<sup>14</sup> This expert system allegedly helped DEC save about 40 million dollars per year by automated decision-making. Encouraged by this success, many countries, including Japan and the United States, invested heavily in the development of the so-called fifth-generation computers, also referred to as ‘artificial intelligence computers’. However, an apparent drawback was their incapacity to learn algorithmically from data and to address uncertainty in reasoning. Moreover, the high maintenance costs of expert systems and the emergence of less expensive and faster desktop computers pioneered by Apple and IBM directly caused a collapse in the market for such systems, taking AI into the second winter with little apparent hope of re-emerging into the mainstream.

Despite its withdrawal from the public eye and a corresponding reduction in funding, work on such matters did not cease entirely. Developments focused on enhancing the statistical validity of the reasoning produced by AI models. A



**Figure 1.** Categories of AI-related techniques.



**Figure 2.** A brief history of AI.

new paradigm, machine learning (Figure 1), placed considerable emphasis on learning actionable insights from complex data and generated excitement within the wider scientific community. New key algorithms and methods were intro-

duced, including expectation-maximization,<sup>15</sup> Bayesian networks,<sup>16</sup> support vector machines,<sup>17</sup> and decision trees.<sup>18</sup> Instead of being explicitly “programmed”, as was the case for the expert systems, machine learning models are “trained” to

discover patterns in data. An example would be in predicting which molecular features of a group of chemical structures is associated with a particular panel of biological effects and making inferences from such descriptions into unknown territory, such as whether a molecule is likely to display considerable toxicity. In a way, this concept enables the automatization of much of the work conducted in the field of quantitative structure–activity relationship (QSAR) modeling in cheminformatics.

The current AI boom started in the late 20th and early 21st centuries, driven by the rapid growth of stored data (“big data”), a concomitant increase in computing power (graphics processing units (GPUs), Google’s tensor processing units (TPUs), etc.), and the continuous optimization of machine learning algorithms (e.g., deep learning).<sup>19</sup> This confluence of events has been revolutionary: for the first time, we can train nontrivial combinations of network elements, in a reasonable time frame, on large amounts of data, simultaneously increasing the practicality of such models. In particular, the feasibility of training deep hierarchical network models in reasonable time has demonstrated the phenomenal capabilities of such approaches in areas as diverse as e-commerce, gaming (e.g., AlphaGo,<sup>20</sup> Poker,<sup>21</sup> and DOTA2<sup>22</sup>), medical image analysis, and self-driving vehicles.

## 1.2. Brief Introduction to the History of Artificial Intelligence Application in Drug Discovery

While AI is not new, neither is the application of AI to drug discovery, especially in modeling generalized structure–activity relationships. Actually, the idea of taking experimental values and using a “descriptor set” for regression goes all the way back to (and possibly even before) Hammett’s pioneering formula connecting reaction rates and equilibrium constants for reactions of benzene derivatives<sup>23</sup> and the computer-assisted identification and quantification of physicochemical properties of bioactive molecules by Hansch,<sup>24,25</sup> who is commonly considered the “father of QSAR” as practiced in the pharmaceutical industry. An increasing number of medicinal chemists have applied various AI methods ever since, to address the central challenge of evaluating and predicting the biological effects of chemicals.<sup>26,27</sup> One method worthy of special mention is the pattern recognition approach,<sup>28–32</sup> which focuses on the elucidation and examination of patterns shared between chemical entities, relying on the general assumption that compounds with similar structural patterns should have similar physicochemical properties and *in vitro* biochemical effects.<sup>33–35</sup> Early prototypes and implementations of neural networks (e.g., the Perceptron<sup>36</sup> and its improved derivatives) emerged, which had potential as a means of solving such problems. Neural networks began to have an impact in the pharmaceutical industry around 1990,<sup>37–43</sup> due to their utility as engines for pattern recognition. A typical example from this time is an article by Weinstein et al.<sup>44</sup> from 1992, in which neural networks were developed to predict the mechanisms of action in a cancer drug screening program. In 1994, the first fully automated molecular design method based on neural networks and evolutionary algorithms was published.<sup>45,46</sup> These integrated learning and decision-making models represent the first instances of constructive machine learning and possess all aspects of AI, namely, the ability (i) to solve problems, (ii) to learn from experience, and (iii) to cope with new situations.<sup>47</sup>

In addition to the methods mentioned above, a variety of machine learning algorithms were developed [such as the support vector machine (SVM)<sup>48</sup> and random forest (RF)<sup>49</sup>] and applied to drug design to help bridge the gap between serendipity-based and rational drug design. As with all such models, they suffer from the “garbage in, garbage out” (GIGO) issue, and all, in one form or another, struggle with the most consistent challenge, namely, which molecular features should be combined to capture the information that will make for the most accurate predictions.

More capable approaches were called for, bolstering the development of contemporary deep learning, which emerged as a more concrete concept around 2010, after isolated developments in the 1990s.<sup>50,51</sup> The ability of some deep learning methods to explore and predict complex relationships between molecular representations and observations (bioactivity, toxicity, etc.) provides a strong basis for optimism that these tools could produce more useful, generalizable insights. Such methods won the Kaggle challenge for compound activity prediction in 2012<sup>52</sup> and the NIH Tox21 challenge for toxicity prediction in 2014,<sup>53</sup> both of which were intentionally difficult challenges, and these methods performed as well or better than the then-current state-of-the-art methods. Pharmaceutical companies that have previously watched modern AI from the sidelines are now jumping in, with several large-scale collaborations between leading pharma and AI companies announced in recent years.

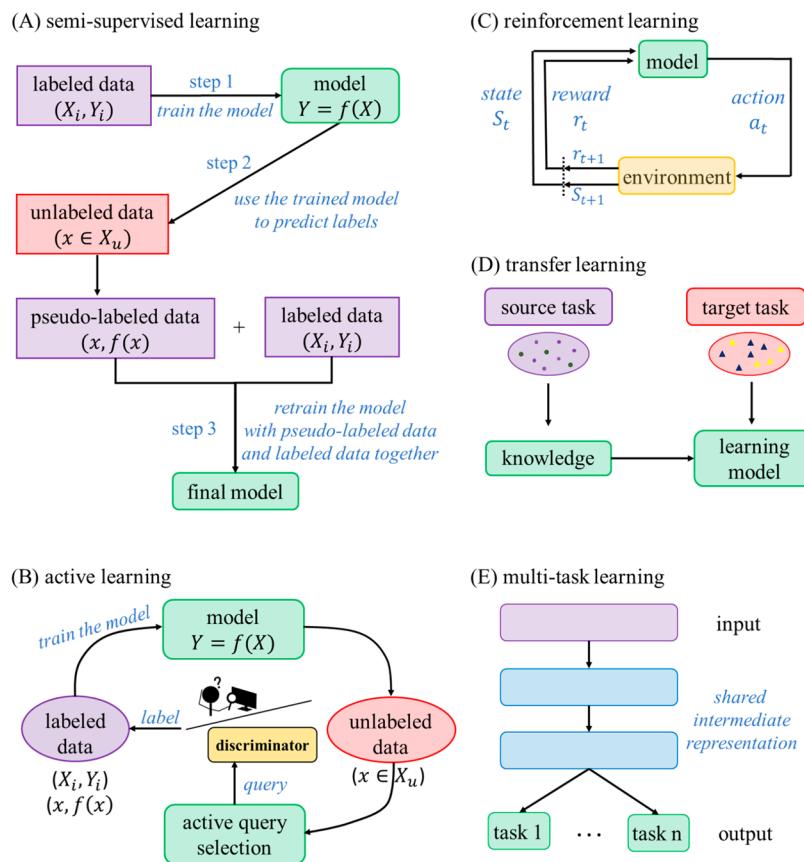
## 1.3. Goal, Scope, And Organization of This Review

The current review differentiates itself from other reviews on related subjects<sup>54–58</sup> by its focus on the following aspects of applied AI in drug discovery. (i) It provides a brief (re)introduction to the mathematical principles underlying the common AI approaches employed in small-molecular drug discovery today, with special attention paid to the advantages and limitations of innovative algorithms. (ii) It provides guidelines for constructing an AI model in medicinal chemistry and an analysis of those factors that impact model performance. (iii) It covers both the achievements and challenges faced by established AI-related techniques for small-molecule drug discovery. (iv) It attempts to look ahead to new possibilities of AI-assisted drug discovery through integration with other disciplines, including the various omics fields and chemistry. Particularly, we point to the closing gap between AI techniques and conventional theoretical chemical methods (quantum mechanics, molecular mechanics) in drug discovery.

The remainder of this review is structured as follows. Section 2 introduces the theoretical aspects of certain popular AI algorithms (section 2.1) and the workflow of constructing an AI model (section 2.2). Considering the vital role of the initial data set for a successful AI model, some guidelines for input data preparation are given. The review continues with a comprehensive summary and analysis of AI techniques used to date for structure-based virtual screening (section 3), ligand-based virtual screening (section 4), de novo drug design (section 5), pharmacokinetics prediction (section 6), drug repurposing (section 7), and the integration of AI into drug discovery (section 8). This review ends with a critical outlook (section 9). Although we have made substantial efforts to provide a comprehensive study of the field as it stands, we will certainly have missed a few pertinent approaches and articles, in part due to the rapid development of the field, and we wish to apologize for any unintentional omission.

**Table 1. Comparison of Different Learning Tasks and Techniques**

learning algorithm	benefits/features	challenges/limitations
supervised learning	the goal is to reconstruct the unknown function $f$ that assigns output values $y$ to data points $x$ assumes a fixed distribution from which the data points are drawn allows clear definition of input and output spaces	requires labeled training data prone to overfitting; there is a trade-off between the simplicity of algorithms and accuracy on the training examples regardless of whether regression or classification modalities are chosen, one requires high-quality data, present in sufficient quantities to enable learning of the underlying mapping
unsupervised learning	many mature and diverse learning algorithms are available can be treated as a geometric or topological problem, that is, the goal is to find similarities and differences between data points, which can be used to spatially order data can be a useful preprocessor for supervised learning and has a chance to discover patterns which have not been previously considered.	cannot generally specify the output space, unlike with supervised methods
semisupervised learning	mix of supervised and unsupervised learning, where less expensive and more abundant unlabeled data can be utilized to train a classifier the main idea of semisupervised learning is to exploit the structural information of unlabeled samples in the feature space to model the distribution of the classes and to find a more accurate classification rule than might be accomplished using only labeled samples	may fail to find useful outcomes from the set of features selected for training since there is no gold standard (such as an outcome variable) and no single objective (such as test set accuracy) if the class labels of the unlabeled samples are automatically determined without verification, the labeling process may introduce a certain number of mislabeled samples into the training set most semisupervised learning techniques intuitively assume that all unlabeled samples come from the same domain as the target set so that the class labels of the target set might be applied to unlabeled samples; in practice, this assumption is too strong for applications involving samples from relevant, but nonidentical, domains
active learning	the learning process iteratively queries the labels of the samples that are expected to be the most informative for an effective training of the classifier, which is expected to improve the learning of the classifier and to reduce the cost and the time of the training sample collection	due to the importance of the parameter values to the final results, model selection is critical for semisupervised learning techniques, which makes it difficult to assess the quality of the obtained solution at convergence
reinforcement learning	an alternative approach to address ill-posed problems when the semisupervised learning assumption does not hold does not require a well-defined initial data distribution; reinforcement learning does not rely on correct input-output pairs; the goal is not to predict the output values $y$ for data points $x$ but instead to find a single value $x^*$ that yields maximum reward $R(x^*)$ . learning can occur in sparse environments (e.g., few or no examples provided, no model of environment, no utility function) has a temporal dimension, i.e., the decision based on the current input will determine the next input	it assumes a reward function that should capture exactly what one needs the system (or machine) to do; usually, this is either a given or hand-tuned offline and kept fixed over the course of learning (there are exceptions, such as inverse reinforcement learning; however, a good reward function is difficult to design; even given a good reward function, local optima may be encountered, hindering global optimization)
transfer learning	especially useful in cases where the system or process is too complex (or too physically hazardous) for teaching machines through trial and error	sample inefficiency for acceptable performance generalization issues involves a conflict between exploitation (of current knowledge) and exploration (of uncharted territory)
multitask learning	requires two specific domains (source domain and target domain) involves training on the source domain and then leveraging knowledge extracted from source domain to speed up learning on the target domain differs from single-task learning in the training (induction) process, where multiple prediction tasks are performed simultaneously to capture intrinsic relatedness tasks can have different data/features	instability in the final results given task A and task B, it can be very difficult to predict whether the knowledge acquired in A transfers to B, often requiring experimentation to determine whether or not it is "appropriate"



**Figure 3.** An illustration of different machine learning processes. (A) Semisupervised learning is a class of machine learning processes that utilizes both labeled and unlabeled data for training. (B) Active learning is a specialized version of semisupervised learning in which a learning algorithm can interactively query the user (or some other sources of information) to determine labels for subsets of the unlabeled data from those regions of the input space where the model is least certain of the correct labeling. (C) Reinforcement learning is a machine learning technique that enables an agent to find the optimal set of actions to promote some outcome, using feedback from its own actions and experiences. (D) Transfer learning methods learn and transfer informative knowledge from old data domains (sources) to a new data domain (target), which can improve the predictive performance on the target domain. (E) Multitask learning models solve multiple tasks simultaneously.

## 2. HOW ARTIFICIAL INTELLIGENCE WORKS

### 2.1. Basic Principles

AI algorithms are a broad field, encompassing a large variety of methods (Figure 1). To help readers better understand AI-assisted drug discovery, we will provide a brief overview of the general principles of those AI-related algorithms that are widely used in drug discovery. Novice readers, or those interested in a broader perspective on the topic, are referred to several excellent monographs<sup>59–61</sup> for a more comprehensive coverage of the fundamentals. Readers with a solid theoretical background can proceed directly to section 2.2.

**2.1.1. Classes of Learning Tasks and Techniques.** The majority of learning tasks and techniques used in drug discovery projects fall into seven classes: supervised learning, unsupervised learning, semisupervised learning, active learning, reinforcement learning, transfer learning, and multitask learning. Each class has its own characteristic benefits and limitations, which are summarized in Table 1. The choice of different techniques should be based on the task at hand. A brief description of these approaches is given below, and Figure 3 schematically illustrates the basic learning processes and principles, respectively.

**2.1.1.1. Supervised Learning.** Many AI learning algorithms use a supervised learning process,<sup>62</sup> in which a set of input data (typically a vector,  $X$ ) and known responses to the data

(output, also called “label” or “target”;  $Y$ ) are required, and the goal of the training process is to learn a mapping function from the input to the output:  $Y = f(X)$ , such that the class labels  $Y$  (in case of a classifier) or target values  $Y$  (for quantitative output data) for unseen data  $X$  can be correctly predicted. For example, given a library of compounds in which each molecule has been labeled as active or inactive, a supervised algorithm can be used to learn the relationship between molecular features and bioactivity, such that new molecules can be predicted to be active or inactive. In general, all classification and regression algorithms may be considered supervised learning approaches, including logistic regression,<sup>63</sup> RF,<sup>64</sup> naive Bayes,<sup>65</sup> and SVM,<sup>66</sup> to name a few popular methods.

**2.1.1.2. Unsupervised Learning.** Unsupervised learning algorithms are used when there are only input data ( $X$ ), with no corresponding output variables. In other words, in an unsupervised model, the data is unlabeled. In such a case, it is difficult to define a useful metric of performance for the algorithm. Instead, the process will extract structures (= features) from the data (= patterns), which can be used to cluster the input samples into different groups. Other tasks involving an unsupervised learning algorithm include density estimation and dimensionality reduction, for instance. The main difference between unsupervised learning and supervised learning is that, in unsupervised learning, no feedback signal is

used to evaluate the quality of the potential solutions. Some popular examples of unsupervised learning algorithms are clustering and projection algorithms (e.g., *k*-means clustering,<sup>67</sup> hierarchical clustering,<sup>68</sup> principal component analysis,<sup>69</sup> and self-organizing map<sup>70</sup>).

**2.1.1.3. Semisupervised Learning.** Semisupervised learning (Figure 3A) sits at a junction between supervised and unsupervised learning approaches<sup>71</sup> and can be useful when many input data ( $X$ ) are available, with only relatively few labeled samples ( $Y$ ). Many real-world drug discovery problems fall into this area. Semisupervised learning can maximize the use of unlabeled data to either modify or reprioritize hypotheses obtained from limited labeled data alone. This feat is usually achieved by (i) using a supervised learning algorithm to train a model with the available labeled data; (ii) applying this trained model to predict labels for the unlabeled data; and (iii) retraining the model with those pseudo-labeled samples and the labeled data. In this way, distributions of the original labeled samples are used for model building and for potentially increasing its predictive power with little additional real-world, e.g., practical experimental, cost.

There are many semisupervised learning methods, and they make different assumptions about the relationship between the unlabeled data distribution and the function learned from the labeled set. Some often-used methods include: self-training, which assumes that a learner's high confidence predictions are correct;<sup>72</sup> cotraining, with the idea to include different "views" of the objects to be predicted;<sup>73</sup> transductive support vector machines,<sup>74</sup> which implement the idea of transductive learning by including unlabeled data in the computation of the margin; and graph-based methods,<sup>75</sup> which rely on the idea of building a graph on the labeled and unlabeled data where instances connected by heavily weighted edges tend to be assigned the same label. Such assumptions are equivalent to prior domain knowledge, and the success of semisupervised learning depends to a large extent on choosing a priori assumptions (constraints, hypotheses) that fit the underlying problem structure well.

**2.1.1.4. Active Learning.** Active learning (Figure 3B) is a specialized version of semisupervised learning,<sup>76,77</sup> which addresses the issue of insufficient labeled training data in a different way. Here, a learning algorithm can interactively query the user (or some other source of information) to determine labels for unlabeled data in the regions of the input space about which the model is least certain. While conventional semisupervised approaches attempt to exploit the latent structure of unlabeled data with the specific goal of improving label predictions, the goal of active learning is to reduce the number of labeled examples needed for learning at the same time.<sup>78</sup> As an active research area, there are many different active learning methods, such as pool-based active learning,<sup>79</sup> where the learner chooses which sample to label next in a pool of unlabeled data; selective sampling,<sup>80,81</sup> where unlabeled data come as a stream and the learner decides to query or discard each arriving point; and batch-mode active learning,<sup>82</sup> where the learner queries multiple instances at once. Interested readers may refer to a report by Settles<sup>83</sup> to learn more about the details of the query strategy frameworks. Active learning strategies combined with other AI algorithms have been shown to match intuitively with the properties desired in many aspects of drug discovery.<sup>76,77,84–86</sup> These strategies include active learning with support vector machine models to identify active thrombin ligands<sup>87</sup> and to classify

compounds into multiple classes<sup>88</sup> and active learning with RF models for rapid screening of Abl Kinase inhibitors<sup>89</sup> and G-protein-coupled receptor modulators.<sup>90</sup> Essentially, they offer a means of "filling in the gaps" in our knowledge of the relationship between the chemical and the biological spaces.

**2.1.1.5. Reinforcement Learning.** Reinforcement learning to some extent strives to emulate reward-driven learning.<sup>91</sup> In short, and in its simplest configuration, an agent attempts to find the optimal set of actions to promote some outcome. This goal is accomplished through a combination of analyzing the environment, performing actions to alter that environment (using policies to translate the internal state of the agent into actions), and scoring the outcome of those actions (reward) (Figure 3C). This pattern of rewards and effective penalties allows us to design and optimize systems without necessarily knowing the "correct", or optimal, approach, as long as success under the reward function correlates reasonably well with success at the objective. Typical reinforcement learning algorithms include the Q-learning methods (such as the deep Q-network (DQN)<sup>92</sup> proposed by Mnih et al.), which calculate the gradient of the loss of the action-value function (referred to as the Q-value function) approximator via backpropagation,<sup>92–94</sup> and the policy gradient methods<sup>95</sup> (e.g., asynchronous advantage actor-critic (A3C),<sup>96</sup> trust region policy optimization (TRPO)<sup>97</sup>), which stochastically sample actions from the current policy and then reinforce those that perform well via stochastic gradient ascent. Evolutionary strategies have also been used for reinforcement learning,<sup>98</sup> which do not calculate gradients analytically but by approximation of the gradient of the reward function directly in the parameter space. Interested readers are referred to a comprehensive book by Sutton et al.<sup>91</sup> to learn more about the background and application details.

**2.1.1.6. Transfer Learning.** Transfer learning (Figure 3D) describes a family of algorithms that relax the common assumption that the training and test data should be in the same feature space and follow the same distribution.<sup>99,100</sup> As the name suggests, transfer learning methods learn and transfer informative knowledge from old data domains (sources) to a new data domain (target), which, if successful, would improve the predictive performance on the target domain. Approaches to transfer learning can be categorized into four groups according to the type of knowledge that is transferred from source to target domain: (i) instance-transfer,<sup>100,101</sup> which assumes that certain instances in the source domain can be reused for learning in the target domain; (ii) feature-representation-transfer,<sup>102,103</sup> whose purpose is to learn a useful feature representation that can be utilized in the target domain; (iii) parameter-transfer,<sup>104–106</sup> which assumes that the source and target tasks share certain parameters or prior distributions of the hyperparameters of the models; and (iv) relational knowledge-transfer,<sup>107</sup> an extension of the zero-shot learning approaches that attempt to predict labels for test data even in cases where there are no examples of that label in the training data, by utilizing an estimation of the similarity of the concepts described by the labels themselves. The interested reader is referred to several sources for further details.<sup>99,100</sup> Generally, the ability to transfer knowledge from a source to a target depends on their degree of relatedness.<sup>99</sup> "Brute force transfer" occurs when the source domain is not related or only weakly related to the target. It can lead to performance degeneration and even break the already learned mapping. This is referred to as "negative transfer". One strategy to

reduce this effect is to transfer knowledge from multiple source domains, e.g., by boosting.<sup>108</sup>

**2.1.1.7. Multitask Learning.** Another learning technique, multitask learning (Figure 3E), has also gained popularity in recent years in the field of drug discovery.<sup>52</sup> Instead of learning only one task at a time, as in single-task learning, several different but conceptually related tasks (that share the same set of features) are learned in parallel and make use of a shared internal representation.<sup>109</sup> This concept is particularly attractive for inferring multitar get prediction models. Although multitask and transfer learning are closely related, they differ in their treatment of the tasks to be managed, with the assumption of equality of tasks in multitask learning, in contrast to a well-defined asymmetry in transfer learning.

**2.1.2. Learning Algorithms.** In this subsection, we recall the mathematical principles of machine learning algorithms that are currently popular in drug discovery. We have chosen to group them by a broad definition according to functional similarity. We note that there are several algorithms that could fit into multiple groups, such as the RF approach, which implements both a decision-tree inspired algorithm and an ensemble algorithm. This compilation does not aim to be exhaustive with regard to either the groups defined or the algorithms chosen; it is rather meant to guide the reader in obtaining an idea of the lay of the land.

**2.1.2.1. Bayesian Algorithms.** Bayesian methods are those that explicitly apply Bayes' theorem to classification and regression problems.<sup>110–112</sup> Suppose that we intend to predict the class  $C$  of an observation based on a vector of features  $(F_1, F_2, \dots, F_k)$  for a given data set. We can do so by considering the posterior probability of each class given our data, through:

$$\operatorname{argmax}_c p(C = c | F_1 = f_1, \dots, F_k = f_k) \quad (1)$$

Since the value of the posterior probability  $p(C = c | F_1 = f_1, \dots, F_k = f_k)$  is difficult to calculate in practice, it is usually transformed by the Bayes' theorem:

$$\begin{aligned} p(C = c | F_1 = f_1, \dots, F_k = f_k) \\ = \frac{p(F_1 = f_1, \dots, F_k = f_k | C = c)p(C = c)}{p(F_1 = f_1, \dots, F_k = f_k)} \end{aligned} \quad (2)$$

where the denominator can be considered constant. Here we explicitly condition a posterior likelihood on our a priori likelihood of the class itself and on the distribution of the data we wish to label. Although somewhat simpler, the accurate estimation of posterior probability remains intractable due to the difficulty of evaluating the coconditioned distributions via the chain rule, in another example of the curse of dimensionality. To address this problem, a strong (naïve) assumption that the features are mutually independent is employed:

$$p(F_1 = f_1, \dots, F_k = f_k | C = c) \approx \prod_{i=1}^k p(F_i = f_i | C = c) \quad (3)$$

Substituting the above formula into the Bayes classifier results in a *naïve Bayes* classifier:

$$\begin{aligned} & \operatorname{argmax}_c p(C = c | F_1 = f_1, \dots, F_k = f_k) \\ & \propto \operatorname{argmax}_c p(C = c) \prod_{i=1}^k p(F_i = f_i | C = c) \end{aligned} \quad (4)$$

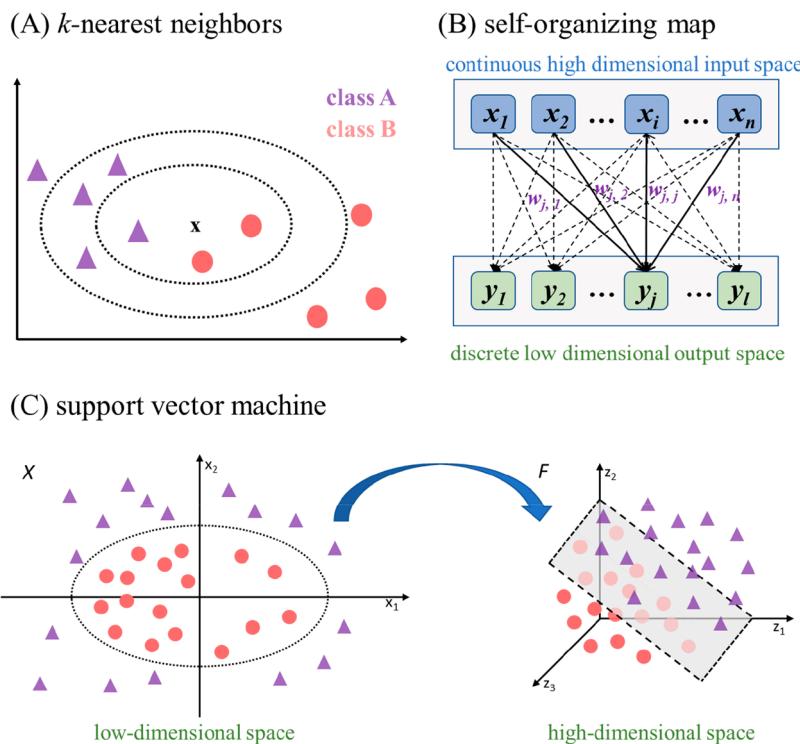
Although it makes the strong, i.e., naïve, assumption that all of the features are class-conditionally independent, the real-world performance of naïve Bayes-based classifiers is often acceptable.<sup>113</sup> However, due to the same assumption, there are some problems that naïve Bayes classifiers cannot solve, such as the well-known XOR problem.<sup>114–116</sup> In these cases, one can instead choose a Bayesian network that does not assume the conditional independence of features.

A Bayesian network, also known as belief network,<sup>117–119</sup> are graphical models, specifically a directed acyclic graph, in which the nodes are variables (in computer-aided drug design, these nodes might refer to molecular descriptors, activity data, or incidental information about an assay), and the edge weights are conditional dependencies. Bayesian networks are able to model many types of relationships (e.g., linear, nonlinear, combinatorial, stochastic) between the variables, despite the comparably high computational cost of training a large Bayesian network with thousands of free variables.<sup>120,121</sup> Due to their probabilistic nature, Bayesian networks are capable of handling noisy or ambiguous data to some extent. However, challenges remain when applying Bayesian networks to limited experimental data, e.g., from biomedical or biochemical experiments.<sup>122</sup> Interested readers are referred to an authoritative article by Friedman et al.<sup>110</sup>

Here, we would like to note the difference between a Bayesian network and other popular artificial neural networks (for more details of artificial neural network algorithms, see section 2.1.2.6). The main differences between these two models lie in their components and in the composition of their graph structure. Unlike the explicit relations between the data features codified in a Bayesian network model, the nodes and edges of an artificial neural network do not necessarily describe the relation between "real-world" data but simply accept input and convert this signal into output values using a series of basic mathematical operations. Bayesian networks allow one to consider the impact of individual components in a simple fashion, and have an advantage over non-naïve Bayesian models in that they display the local Markov property,<sup>123,124</sup> informally that each node is conditionally independent from each of its ancestors apart from its immediate parents. In practice, this means that we do not have to consider the coconditionality of each pair of variables, which markedly decreases the computational complexity of estimating the probability distributions for a model.

**2.1.2.2. Instance-Based Methods.** Instance-based (sometimes called "memory-based") methods<sup>125</sup> are a family of "lazy" learning algorithms in which all training examples  $\langle x_i, f(x_i) \rangle$  are stored and generalization is postponed until a new query instance  $x_q$  is classified. The target function  $\hat{f}(x_q)$  is approximated directly from the training instances through the analysis of the relationship between  $x_q$  and instances seen in the training set.

The most basic example of instance-based methods is the  $k$ -nearest neighbor ( $k$ NN) method<sup>126,127</sup> which assumes that each instance corresponds to a point in an  $n$ -dimensional space  $\mathbb{R}^n$ . Given a query instance  $x_q$ , its  $k$ -nearest neighbors (Figure 4A) are defined as the data points with the  $k$  smallest distances



**Figure 4.** Schematic representations of (A) *k*-nearest neighbors, (B) self-organizing map, and (C) support vector machine algorithms.

to  $x_q$ . Then, the target function  $\hat{f}(x_q)$  is evaluated for the nearest neighbor list. A *k*NN method either takes a vote among its *k*-nearest neighbors for a discrete-valued target function or computes the mean of  $f$  values of *k* nearest neighbors for real-valued target functions ( $\hat{f}(x_q) \leftarrow \sum_{i=1}^k f(x_i)/k$ ). In *k*NN methods, the weighted-average smooths out the influence of isolated noisy training instances. However, the assumption that the classification of  $x_q$  will be similar to the classification of other instances that are nearby, according to their distance in the *d*-dimensional feature space, can cause problems for *k*NN methods. This is a fundamental issue for high-dimensional distance comparisons in general,<sup>128</sup> due to two major issues. First, the odds of a separation-relevant signal being swamped by noise scales with the number of features. That is, when instances are described by many attributes (for example, more than 20 attributes), but only few (for example, only 2) are relevant to the target function, *k*NN methods will be easily misled by irrelevant attributes in high-dimensional space. Second, distance metrics following the form of the Manhattan and Euclidean metrics have a limited ability to separate the maximally and minimally distant points for every point in the input space. Solutions to this problem include weighting features and removing less-relevant attributes from the instance space, e.g., by optimizing feature scaling with an evolutionary algorithm<sup>129</sup> or performing feature extraction and dimension reduction prior to applying *k*NN algorithms.<sup>130–132</sup> A drawback of the *k*NN method is that the predictions must be made within the activity interval of the training data. Furthermore, a *k*NN classifier has an increasing probability of making a mistake as it approaches the boundary.<sup>133–136</sup> Considering the fact that the knowledge of boundary points can facilitate data mining, as they represent a subset of the population that possibly straddles two or more classes, several methods<sup>137–140</sup> have been proposed to find boundary points

efficiently and effectively in multidimensional data sets and increase the classification accuracy.

Another group of instance-based methods are radial basis function networks (RBFs).<sup>141,142</sup> This learning technique is closely related to distance-weighted regression but “eager” instead of “lazy”; that is, it generalizes before a new query instance is provided based on the following equation:

$$\hat{f}(x_q) = \sum_{i=1}^N w_i \cdot \phi(\|x_q - c_i\|) \quad (5)$$

where  $N$  is the number of neurons in the network,  $w_i$  is the weight of a given neuron,  $c_i$  is the corresponding center vector, and  $\phi$  is the basis function employed. One common choice of basis function is a Gaussian function centered at  $c_i$ .

The function of  $\hat{f}(x_q)$  can be viewed as describing a two-layer neural network (for more information on neural networks, see section 2.1.2.6) where the first layer of units computes the various values gained from evaluation of the basis functions and the second layer creates a linear combination of the results. It is generally accepted that the choice of nonlinear activation function for the hidden nodes is not critical for performance,<sup>143–145</sup> but the performance of an RBF network depends upon the locations of the RBF neuron centers.<sup>146,147</sup> In practice, the centers (means) can be restricted to be a subset of the training data or can be learned by unsupervised training procedures, including clustering,<sup>142</sup> supervised vector quantization,<sup>148</sup> and classification tree algorithms.<sup>149,150</sup> Overall, an RBF-network can model non-linear separating hypersurfaces, with an overall network complexity similar to the multilayer perceptron models (for more information, see section 2.1.2.6).

The self-organizing map (SOM) is another example of the instance-based methods.<sup>70</sup> The SOM is a special class of neural

networks based on competitive learning, as contrasted with error-correction learning in other neural networks. The training process itself takes unlabeled data and maps them from a continuous, high-dimensional input space (e.g., a multidimensional descriptor space) to a discrete, low-dimensional output space specified by the user. A particularly useful feature of this method is that, when employed with a sufficiently large number of nodes, it can preserve topological characteristics of the input space, i.e., input data that are close together in the input space will also be close in the output space (Figure 4B). The structure of a SOM is simple, yet quite powerful. The algorithm tessellates the input space (for example, chemical space defined by molecular descriptors) into a predefined number of clusters. Each cluster is represented by a centroid (“neuron”) in the SOM network. All neurons are placed at the nodes of an  $n$ -dimensional lattice (the output space), where  $n$  is typically in the range of one-to-three. Higher-dimensional maps are also possible but uncommon in drug discovery<sup>151,152</sup> as it impedes their utilization as visualization tools without necessarily increasing the descriptive power of the model. Each neuron in the lattice is fully connected to all of the source nodes in the input space. During the competitive learning process, the algorithm identifies the best-matching unit for each new training example, by finding the neuron with the closest corresponding weight vector to that data point. Following this, the best-matching unit and nearby neurons update their weight vectors to minimize the distance to the new training data. The magnitude of the allowed-change, and the manner in which neighboring neurons of the best-matching unit are chosen, can be used to specify the manner in which the quantization error is reduced. In addition, hard constraints can be placed on topographic error, forcing the SOM to maintain the valuable neighborhood property at the expense of some increase in quantization error. With a sufficiently large number of neurons, the SOM trains in such a way that a topographic map of the input data is generated on the lattice, where the spatial locations (i.e., coordinates) of the neurons indicate the inherent statistical features in the input data. In general, the SOM is suitable for data survey because it provides a natural visual overview of the data space. However, there are still some difficulties in its utilization, even given the large body of published research featuring applications of the SOM over the years. First, determining the optimal network architecture remains far from trivial for instances where the data are poorly characterized.<sup>153</sup> Second, hierarchical relations between input data cannot be addressed conveniently within the framework of the SOM, and their proper identification remains an important data mining task.<sup>154</sup> Finally, both the winner-search and updating processes are time-consuming operations for very large maps.<sup>155</sup> This problem can be alleviated by obtaining a good estimate of the initial values of the models. Interested readers are referred to refs 156 and 157 for more hints on the construction of very large SOMs.

Another instance-based method is the SVM,<sup>17</sup> which can perform both classification and regression tasks.<sup>158</sup> The idea of the SVM is that training data that are nonlinearly separable in their original, low-dimensional input space can be separated in a higher-dimensional latent space, which is constructed through mapping (Figure 4C). SVM methods normally incorporate this mapping via the “kernel trick”, which is essentially a means of reducing the computational overhead of the process. By utilizing this feature mapping to avoid explicit

calculation, the SVM fits a nonlinear separation in the original space without ever actually locating each point in the transformed high-dimensional space. The kernel functions used vary, as do the hyperparameters employed, but are often Gaussian or a related RBF. The fitting process itself consists of finding the hyperplane in the high-dimensional space that maximally separates two classes, in the binary classification setup. The hyperplane can have either hard- or soft-margins, essentially corresponding to the acceptable degree of contamination of the data in the hyperplane region. Soft-margin approaches tend to allow for better generalization, at the cost of increased noise. SVMs scale comparatively well with the number of features in their input space, as a result of their focus on maximal-margin separation.<sup>66</sup> However, in practice, feature selection is normally performed and accomplished using methods that minimize the generalization error of the model,<sup>159</sup> to reduce the impact of unhelpful features and improve time-performance. Multiclass classification is common in practical applications such as disease detection and compound activity recognition. SVMs were originally designed for binary-classification problems, and the optimal means of extending their utility to multiclass problems is still an ongoing research issue. Currently, there are two common types of approaches. One approach is based on partitioning strategies<sup>160</sup> (including one-versus-one,<sup>161</sup> one-versus-rest,<sup>162</sup> rest-versus-one,<sup>163–165</sup> one-versus-one-versus-rest,<sup>166</sup> and other strategies<sup>167</sup>), while the other approach is based on reformulating the construction of the multiclass SVM classifier by solving larger optimization problems.<sup>168</sup> Readers who are interested in a comprehensive discussion of the SVM methods are referred to the literature.<sup>17,169</sup>

**2.1.2.3. Decision Tree Algorithms.** Decision trees (DTs) are tree-shaped representations of chains of decisions that can be used to classify, or predict values for, input data.<sup>170</sup> They are suitable for both categorical (categorical variable decision tree) and continuous (continuous variable decision tree) input and output variables. Given a training data set represented in the root node (Figure 5A), a DT model splits the population of data into two or more subsets (represented as decision nodes). The splitting will continue to divide these into progressively smaller subsets by posing an either-or scenario until an outcome (represented as a terminal node) is encountered. In the naive case, this would be where every leaf of the tree corresponds to one training datum with its associated label, but in practice, it is terminated before this stage, to improve the generalizability, and to reduce the tendency to overfit, of the model. Various decision tree algorithms have been developed, for example, ID3 (iterative dichotomizer 3),<sup>171,172</sup> C4.5,<sup>173</sup> and CART (classification and regression trees).<sup>174</sup>

There are several advantages of DT methods, such as innate support for multiclass problems,<sup>175</sup> flexibility with regard to the input and output data types (numerical or categorical variables), and the legibility of the resulting models, giving a clear and unambiguous rationale for each of its labeling decisions.<sup>176</sup> The major issues with such approaches are poor handling of noisy or incomplete data,<sup>177</sup> the expense of modeling many partitions of the data that do not correlate well with those observed in a linear separation method,<sup>175</sup> and overfitting<sup>178</sup> and sensitivity to the starting state.<sup>175</sup> The overfitting problem usually occurs when a DT model grows many branches, due to outliers and irregularities in the data, which are essentially captured as “special cases” rather than as potential noise. The sensitivity issue is closely related; the final

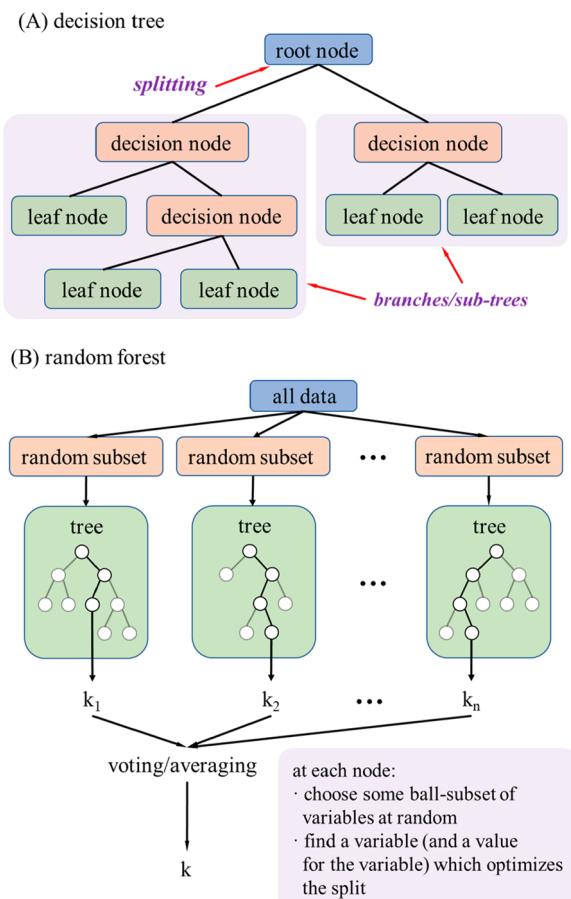


Figure 5. Schematic diagrams of the (A) decision tree and (B) random forest algorithm.

construction of a tree is dependent on its training data set, more so than other approaches, due to its definition of decision boundaries at multiple levels. Two approaches we can use to reduce the impact of this overfitting effect are setting constraints on model parameters (e.g., tree size) and tree pruning (prepruning, postpruning). Pruning involves removing branches that employ features with low overall importance, and this can be accomplished by removing those branches that do not contribute substantially to the overall accuracy of the model. DT methods are not suitable for rigorously predicting continuous numerical variables, although their performance on that task can be improved with variants incorporating more

complex rules, such as linear combinations of features in a given decision path across the tree graph.<sup>179</sup>

Like many other models, tree-based models are plagued by bias and variance. A powerful way to alleviate this problem is to use an ensemble approach<sup>180</sup> (for more information, see section 2.1.2.4) to transform DT into an RF model.<sup>64</sup> As illustrated in Figure 5B, each tree is planted and grown as follows: (i) create multiple pseudorandom data subsets through sampling with replacement from the original data; (ii) specify a number  $m$  that is smaller than the number of input variables  $M$ , and select  $m$  variables at random out of  $M$  to consider when splitting a node; (iii) grow each tree to the largest extent possible without pruning; and (iv) generate new predictions by aggregating the predictions of every tree (i.e., majority votes for classification, average for regression).

In this way, the RF method preserves the advantages of the DT method and potentially overcomes several challenges. For example, it can process thousands of input variables and identify those that are most important, so it can be considered a means to accomplish dimensionality reduction.

**2.1.2.4. Ensemble Algorithms.** Ensemble learning employs the same learning algorithm to train multiple predictive models, improving their accuracy and reliability compared to their single-model instances.<sup>181</sup> It often allows modelers to obtain an idea of the fragility of the model or its dependency on certain data points, which can help when deciding which new data sets should be acquired, and with what priority. Some of the commonly used ensemble algorithms include bagging,<sup>182</sup> boosting,<sup>183</sup> and stacking.<sup>184</sup>

Bagging reduces the variance of a prediction by combining multiple models (learners) trained on different subsamples of the same data set (Figure 6A).<sup>182</sup> The process includes (i) creating multiple data sets drawn from the original data, (ii) training multiple classifiers on each data set, and (iii) combining all models to generate a single response value, e.g., mean, median, or mode, depending on the problem at hand. There are various implementations of bagging models, one of which is the RF approach we have discussed in section 2.1.2.3.

Boosting reduces bias and variance by combining insights from a number of weak learners to form a comparatively strong learner. It is an ensemble technique requiring relatively large data sets.<sup>183</sup> As displayed in Figure 6B, boosting begins with the training of a first learner on a data set through random sampling with replacement from weighted data, followed by the sequential creation of multiple learners that attempt to correct the errors of the previous learners until the training set is perfectly modeled or a maximum number of models is

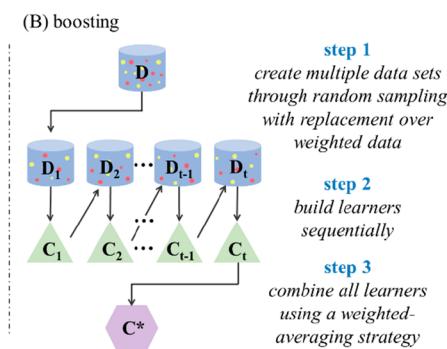


Figure 6. Illustrations of (A) bagging and (B) boosting ensemble algorithms.

reached. The final prediction is an accuracy-weighted average of all learners. Boosting can produce models with a higher accuracy than bagging, with the associated cost of an increased risk of overfitting. Various boosting algorithms have been introduced, including adaptive boosting (AdaBoost),<sup>185</sup> gradient tree boosting,<sup>186</sup> and eXtreme gradient boosting (XGBoost),<sup>187</sup> each of which has multiple parameters to define the search strategy used in finding a good summary predictor.

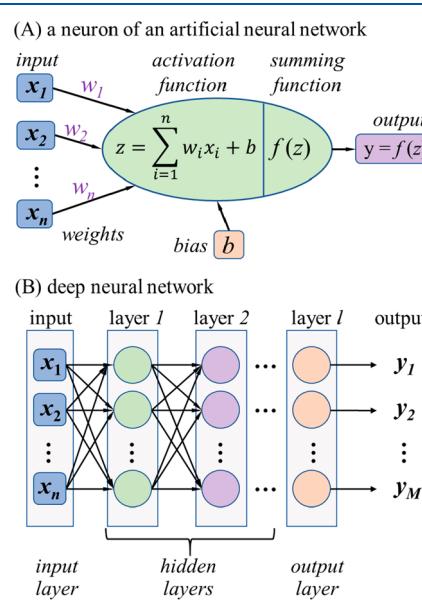
There are several differences between bagging and boosting techniques. (i) The approach to multiple data set generation. Bagging requires that each portion of the training data have the same probability to be used for model training. In contrast, each of the data samples in boosting has an associated weight based on the success with which their labels have been predicted to date, and therefore, some are more frequently involved in the new sets. (ii) The training sequence. In bagging, each model is trained independently. Boosting has a cascade format, in which each learner stands on the shoulders of its direct antecedent and is trained on a new subset, considering the previous learners' successes and failures. (iii) The prediction protocol. In bagging, the output is obtained by averaging the responses of the all learners or by majority voting. Boosting, assigns a weight to each learner and computes a weighted average of their estimates. Furthermore, in some boosting algorithms, an extra condition is involved in postprocessing of the output of each model. For example, in AdaBoost,<sup>185</sup> models having an accuracy of less than 50% simply have their class predictions flipped. Additional model pruning can be applied to both methods post hoc; here, criteria are used to remove under-performing models or those that are essentially homogeneous.

Stacking<sup>184</sup> involves training a series of models on a particular problem, passing the outputs of these to a final, adjudicating model that linearly combines these in a weighted fashion and provides a final prediction for any given input datum. This final function can take any suitable form and largely depends on the type of the ultimate output, but ridge and logistic regression are common choices.<sup>188</sup> This approach has proven power in improving predictions for complex, sparse data sets,<sup>189</sup> and has been employed in the field of chemometrics for the optimization of automated spectral analysis.<sup>190</sup>

**2.1.2.5. Dimensionality Reduction.** The individual variables of a data set are often information-poor or strongly correlated. Dimensionality reduction algorithms reduce the number of variables by mapping the data into a new space with fewer dimensions than the original space. Ideally, the new space retains all of the relevant information, while reducing the space and complexity required to represent it. There are two major approaches to performing dimensionality reduction: (i) selecting a subset of the original feature set (feature selection) and (ii) extracting a new set of features from the original feature set (feature extraction). Both feature extraction and feature selection play an important role in improving learning performance, preventing overfitting, building better generalization models and increasing computational efficiency. Feature extraction is preferred in situations where the original input data do not contain features that a given learning algorithm can immediately understand.<sup>191</sup> Unlike creating a new set of features in feature extraction algorithms, feature selection algorithms aim to find a subset of the most representative features, thereby keeping some of the original features and retaining their corresponding meaning, which

often results in higher model readability and interpretability. Therefore, feature selection algorithms are more popular on large-scale data applications (e.g., genetic analysis). Feature selection techniques tend to have much lower computational complexity than feature extraction algorithms.<sup>192</sup> The various dimensionality reduction algorithms are generally divided into linear and nonlinear algorithms, and there are a variety of linear and nonlinear techniques, including but not limited to principal component analysis (PCA),<sup>193,194</sup> linear discriminant analysis (LDA),<sup>194–196</sup> maximum margin criterion (MMC),<sup>197</sup> locally linear embedding (LLE),<sup>198</sup> ISOMAP,<sup>199</sup> and t-SNE.<sup>200</sup> PCA and LDA have historically been the two most popular dimensionality reduction techniques for feature extraction due to their relative simplicity and proven efficacy. Both of these algorithms assume that the data of each class are Gaussian distributed and extract features by projecting the original data vectors into a new feature space through a transformation matrix. It should be mentioned that this assumption also leads to a key limitation of techniques such as PCA, LDA, and their variants, since data rarely follow a Gaussian distribution in practice and are quite often multimodal, leading to a collapse in the descriptive power of these techniques. The two approaches optimize the transformation matrix with different intentions.<sup>201</sup> As an unsupervised method, PCA describes the directions in the projected space with maximal variance.<sup>193,202,203</sup> LDA is supervised and pursues the projection directions that are most effective in discriminating between output classes.<sup>204–206</sup> t-SNE is an increasingly popular tool for dimensionality reduction, specifically for the purposes of visualization, based on the preservation of neighborhoods identified in the original high-dimensional input space in a lower-dimensional output space.<sup>207</sup> Readers who are interested in more details on the specifics of dimensionality reduction approaches are referred to several overviews.<sup>208–213</sup>

**2.1.2.6. Artificial Neural Networks.** Artificial neural networks (ANNs) are composed of interconnected artificial neurons that act as basic information-processing units.<sup>214–219</sup> Figure 7A displays a neuron (or node) in more detail. The



**Figure 7.** A simple illustration of feed-forward artificial neural networks (ANNs): (A) a neuron in an ANN and (B) architecture of deep neural networks (DNNs).

product of the input vector  $\mathbf{X} = [x_1 \cdots x_n]^T$  and the neuron weights  $w_i$  is combined with a bias term  $b$  and passed through an activation function  $f(z)$  to generate an output. A neuron can be regarded as a function that maps an input vector to an output vector. Mathematically, the output of the neuron can be represented as

$$\mathbf{y} = f\left(\sum_{i=1}^n w_i x_i + b\right) \quad (6)$$

A typical ANN architecture contains many artificial neurons arranged in a series of layers: the input layer, an output layer, i.e., the top layer, which generates a desired prediction (absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, activity, a vector of fingerprint etc.), and one or more hidden (middle) layer(s) where the intermediate representations of the input data are transformed.

A variety of learning techniques can be used to train an artificial neural network, for example, gradient descent algorithms such as stochastic gradient descent and Adam,<sup>220</sup> coupled with backpropagation. Backpropagation refers to the distribution of the calculated error at the output of a model back through its structure, assigning a portion of the total network error to each neuron and enabling the gradient descent algorithm to find a better solution, i.e., the approximation of the underlying function to the problem at hand. Error backpropagation is used primarily for supervised learning tasks, as it requires the target output values for error calculation, although it also works with quantization or construction errors, enabling its employment for architectures such as the autoencoder (section 2.1.2.6.4). For more mathematical details about how the backpropagation algorithm works, we refer readers to several excellent references.<sup>217,221,222</sup>

The term “deep learning” refers to any learning system composed of several information processing layers.<sup>223</sup> These layers might incorporate multiple machine-learning methods, although this term currently refers mainly to one or more variants of artificial neural networks. A “deep neural network” (DNN) (Figure 7B) refers to an ANN that has several hidden layers (with the definition of “deep” being a matter of some debate) with several differences. (i) Unlike traditional ANNs which have typically been used for supervised learning tasks, DNNs can be applied to both supervised and unsupervised tasks. (ii) The workhorse algorithm used for training an ANN, i.e., adjusting its parameters, has been stochastic gradient descent (SGD) coupled with backpropagation.<sup>224</sup> However, this is usually employed in concert with a selection of other algorithms and designs to successfully train a DNN, examples of which include regularization (such as maxout<sup>225</sup> or dropout<sup>226</sup>), rectification of activation functions, and optimization of structures, as in Microsoft’s ResNet approach.<sup>227</sup> In addition, different optimizers, such as Adam<sup>220</sup> and evolutionary methods, e.g., swarm algorithms,<sup>228</sup> can be employed. The extent to which these improve matters is an area of active research and debate in the AI community. (iii) In contrast to using hand-crafted features extracted by extensive preprocessing and feature engineering efforts in traditional ANNs, DNNs can automatically extract useful features from raw input through their hierarchical structure. For example, given a molecular substructure fingerprint representation, for instance, DNNs are capable of learning that a given pattern of bits corresponds to a given feature and that to a certain

biological activity. This ability to abstract information, and to generalize, is behind much of the success of these methods. (iv) DNN training requires a relatively large number of training examples compared to human learning. Creating a system that can generalize from few examples is the focus of ongoing research. To a certain extent, the availability of data limits the tasks we can accomplish today.

ANNs and DNNs are theoretically similar, and therefore, in the following sections, we will discuss them as a single concept.

**2.1.2.6.1. Feed-Forward Network.** The feed-forward neural network is the most basic form of artificial neural network, where the connections within the network architecture are directed from the input to the hidden layer(s) and onward to the output layer, without loops or backward connections. Several examples of feed-forward networks include, but are not limited to, the single-layer Perceptron, multilayer Perceptron, RBF, SOM, and deep feed-forward networks. We refer readers to section 2.1.2.2 for more information about the RBF and SOM approaches.

**2.1.2.6.2. Recurrent Neural Network.** A recurrent neural network (RNN) is an architecture of ANNs designed to identify patterns in sequential data, such as time-series data, genome, and protein sequence data, or simplified molecular input line entry specification (SMILES) strings.<sup>59</sup> In contrast to feed-forward networks, RNNs introduce multiple cells (Figure 8) which take as their input not just the current input

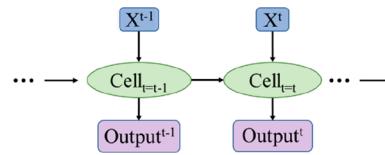


Figure 8. Basic architecture of recurrent neural networks (RNNs).

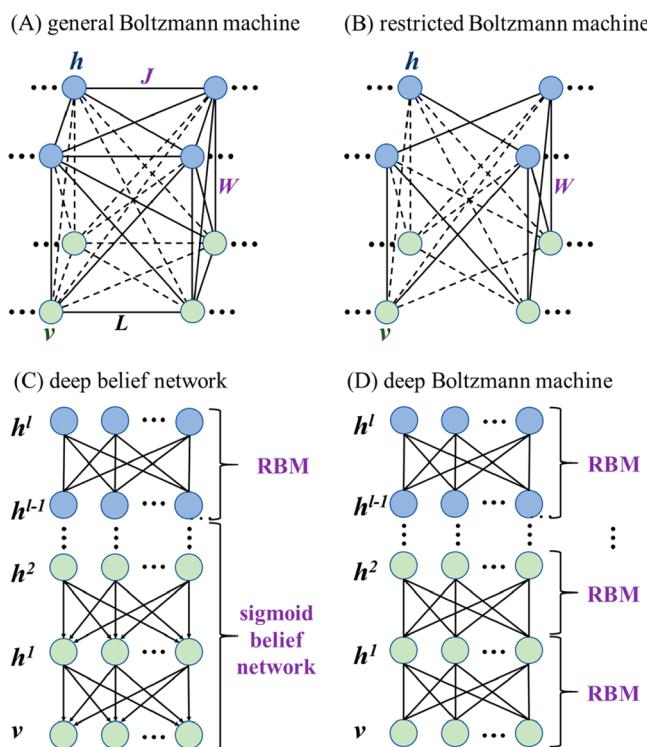
$x^{t-1}$  but also the information they have perceived from the previous state  $cell^{t-1}$ , according to the following equation:

$$cell^t = \sigma(Wx^t + Ucell^{t-1})$$

where  $\sigma$  is the activation function,  $W$  is the weight matrix, and  $U$  is the hidden-state-to-hidden-state matrix (i.e., transition matrix). A regular RNN is unable to capture long-term dependencies because of the vanishing gradient problem.<sup>229</sup> Several variations of RNNs were specifically designed to alleviate this problem, such as long–short-term memory (LSTMs)<sup>230</sup> and gated recurrent unit (GRU).<sup>231</sup>

Deep RNN models based on LSTM or GRU have been used for de novo molecule design. These RNN models<sup>232</sup> are able to capture long-term dependencies and approximate grammars, which are necessary to conduct SMILES string prediction since a valid SMILES string, in addition to the correct valence for all atoms, must count ring openings and closures, as well as bracket sequences with several bracket types. All of these elements combine to form a grammar relating chemical vocabulary to molecular features and should be captured by the model to compute valid SMILES that correctly represent the intended molecule.

**2.1.2.6.3. Boltzmann Machine and Its Variants.** A Boltzmann machine<sup>233</sup> is an energy-based network. It contains a set of visible binary units  $v \in \{0,1\}^V$  that represent the input of the network (e.g., a fingerprint descriptor as a molecular representation) and a set of hidden binary units  $h \in \{0,1\}^H$  that act as feature detectors, forming a densely connected,



**Figure 9.** Schematics of the (A) general Boltzmann machine, (B) restricted Boltzmann machine, (C) deep belief network, and (D) deep Boltzmann machine.

undirected graph. As illustrated in Figure 9A, each neuron is connected with all the other neurons through the energy function:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\mathbf{v}^T L \mathbf{v} - \mathbf{h}^T J \mathbf{h} - \mathbf{v}^T W \mathbf{h} \quad (7)$$

where  $\theta = \{L, J, W\}$  are the model parameters that represent visible-to-visible, hidden-to-hidden, and visible-to-hidden symmetric interaction terms, respectively. The joint distribution of the energy-based probabilistic model is defined as follows:

$$P(\mathbf{v}, \mathbf{h}; \theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{v}, \mathbf{h}; \theta)) \quad (8)$$

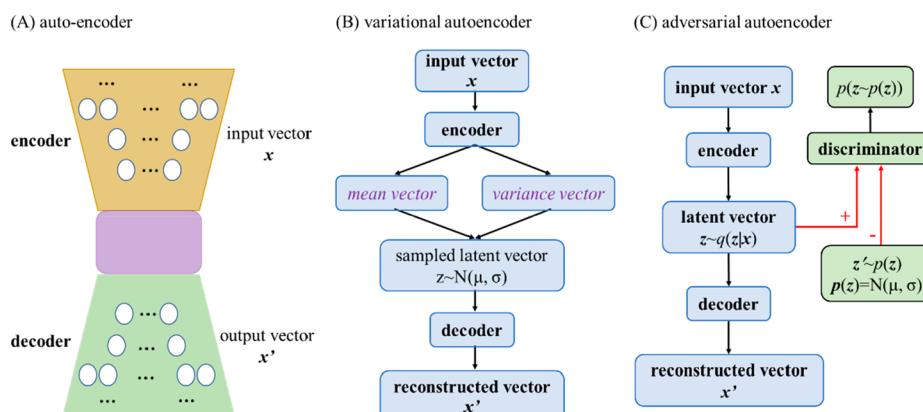
in which the normalization term  $Z(\theta)$  is called the partition function, which is a summation over all pairs of visible and hidden vectors, such that the overall probability distribution sums to unity. One may rarely be interested in training a complicated and fully connected Boltzmann machine because exact maximum likelihood learning with the Boltzmann machine is intractable for nontrivial problems. Several techniques have been introduced to reduce the complexity of this problem, one of which is to restructure the problem as a bipartite graph, in which there are no intragroup connections, i.e., visible layers are only connected to hidden layers, and vice versa. This leads to the “restricted” Boltzmann machine (RBM).

In an RBM model, both  $J$  and  $L$  are set to zero<sup>234</sup> (Figure 9B), which means that no intragroup connections are allowed, substantially reducing the computational complexity of training. The training process for an RBM model employs contrastive divergence with persistent Markov Chain Monte Carlo sampling,<sup>235</sup> although newer techniques for inferring the weights and biases have been developed. RBMs are an efficient way to obtain useful feature detectors. However, their performance can be further improved by stacking several RBMs into either a deep Boltzmann machine (DBM)<sup>236</sup> or a deep belief network (DBN)<sup>237</sup> with a training procedure referred to as greedy layer-wise pretraining.

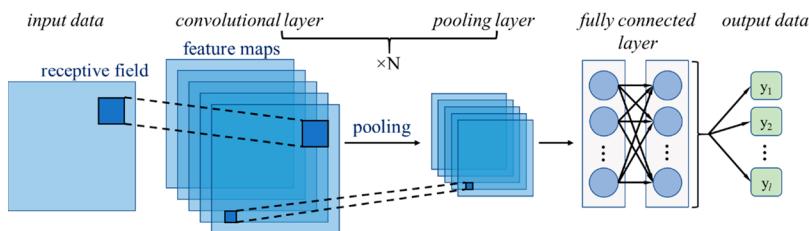
In a DBN (Figure 9C), the connections between layers are directed.<sup>237</sup> The top two layers of the network are modeled as an RBM (an undirected graphical model), and the lower layers form a directed generative model in a downward fashion. For instance, given a DBN with three-hidden layers, the joint distribution is given as

$$P(\mathbf{v}, \mathbf{h}; \theta) = P(\mathbf{v}|\mathbf{h}^1)P(\mathbf{h}^1|\mathbf{h}^2)P(\mathbf{h}^2, \mathbf{h}^3) \quad (9)$$

where  $\mathbf{v}$  is the vector of visible binary units,  $P(\mathbf{v}|\mathbf{h}^1)$  and  $P(\mathbf{h}^1|\mathbf{h}^2)$  are the conditional probabilities, and  $P(\mathbf{h}^2, \mathbf{h}^3)$  is the joint distribution. The learning process of a DBN is through training RBMs one at a time and then stacking them on top of each other to infer successive hidden layers. Once the RBMs are stacked, the prior distribution over the hidden nodes changes, with the result that only the uppermost RBM retains the “correct” graphical model. DBNs are relatively easy to code and can work sufficiently well with just a few layers. In



**Figure 10.** Architectures of (A) autoencoder, (B) variational autoencoder, and (C) adversarial autoencoder.



**Figure 11.** Schematic diagram of a convolutional neural network (CNN).

particular, DBNs can be used to initialize a deep neural network which can be fine-tuned by back-propagation. We refer interested readers to an article from Hinton, Osindero, and Teh.<sup>237</sup>

In a DBM (Figure 9D), the connections between all the layers are undirected and symmetric, such that each pair of layers forms an RBM, each of which is connected.<sup>236</sup> This architecture renders the sampling of the hidden units given the visible units inefficient, and the contrastive divergence method is slow for DBMs. When training a DBM, one encounters two issues. The first is the difficulty of exact maximum likelihood learning due to the hard inference problem induced by the partition function. The other one is that multiple layers of hidden units make learning in DBMs impractical.<sup>238</sup> Aiming to overcome these challenges, a variational approach was proposed by Salakhutdinov and Hinton,<sup>236</sup> in which inference over the states of the hidden variables is carried out by applying variational approaches, such as mean-field variational inference, which is normally the fully factored, fully connected naïve model. Learning can then be performed by using a variational approximation procedure together with Markov chain Monte Carlo sampling to approximate the gradients of the otherwise intractable partition function.

**2.1.2.6.4. Autoencoder and Its Variants.** A minimalist autoencoder (Figure 10A) consists of an encoder functionality, translating an input into a latent space, and a decoder, translating this internal latent-space representation back to the original input space. Given an input  $\mathbf{x}$ , the goal of an autoencoder is to compute a reconstruction  $\mathbf{x}'$  with minimal error compared to the original input  $\mathbf{x}$ , while also having an internal representation with fewer features than are present in the input. In essence, then, an autoencoder can be regarded as a dimensionality reduction approach, as it can be used to build potentially more meaningful, or less noisy, representations of feature vectors. It is also extensively employed in generative modeling, as it allows for the creation of a well-described and consistent internal representation with few dimensions that can be translated back to the original or a related space. When the input data have a complicated internal structure, as is often the case with biological data, multiple autoencoders can be stacked, adding nonlinearity and flexibility. One approach to resolve the increased computational complexity is to adopt a similar approach to that utilized in the deep belief networks (section 2.1.2.6.3), to train in a greedy layer-wise fashion. The advantages and disadvantages of doing so are broadly similar.

Variations of autoencoders have been developed to prevent “vanilla” autoencoders from simply approximating the identity operator and to increase their ability to extract useful features from data. Examples include the denoising autoencoder,<sup>239</sup> the variational autoencoder (VAE),<sup>240,241</sup> and the adversarial autoencoder (AAE).<sup>242</sup> Denoising autoencoders are intended to recover the original undistorted input data from corrupted

input data, making the final model more robust to noisy data than vanilla autoencoders. To accomplish this, stochastic noise is added to  $\mathbf{x}$ , giving  $\mathbf{x}'$ , and the autoencoder is trained as before, with the exception that we are now interested in the loss values obtained from  $\mathbf{x}$  and  $\mathbf{x}'$ , when trained on  $\mathbf{x}'$ .

In the case of a VAE (Figure 10B),<sup>240</sup> we direct the system to produce outputs similar to our inputs by adding a requirement that the distribution of the variables in the latent space should follow some distribution, most commonly a Gaussian. Subsequently, these latent vectors are fed into the decoder to reconstruct the input. Due to this, we have the possibility of generating new samples that are similar to the original samples used for training, which has the effect of increasing the localized density in our latent space. This process has the advantage that input data that are somewhat similar to our training set are more likely to have a sensible decoded form than in simpler autoencoder models. Besides the reconstruction loss used in vanilla autoencoders, another term is also added to the loss function in VAEs, that is, the Kullback–Leibler divergence between the distribution created by the encoder and the prior distribution.

Since the integral of the Kullback–Leibler divergence term in VAEs adds to the computational cost, AAEs<sup>242</sup> (Figure 10C) have been introduced to avoid the necessity of evaluating that term. This is achieved through introducing a new network (called the discriminator) to each AAE, which is trained to determine whether its input comes from encoder-generated latent vectors or from a prior distribution determined by the user. The first step of training the overall network is the same as for the vanilla autoencoder above, where we minimize our reconstruction loss, i.e., we train the encoder-decoder system to reproduce inputs. Then, we train the discriminator to distinguish between noise and true values and predicted true values. Finally, we optimize our encoder by using the discriminator with a cross-entropy loss function; essentially, we use the discriminator to label our input data and drive the optimizer to produce only outputs that are indistinguishable from true data.

**2.1.2.6.5. Convolutional Neural Networks.** Convolutional neural networks (CNNs) are used in situations where input data can be represented as images or image-like objects.<sup>243,244</sup> A typical CNN contains at least three components: convolutional, pooling, and densely connected layers. A convolutional layer is described in terms of its width, height, and depth; that is, it captures  $x$ - and  $y$ -coordinate information over a small receptive field (a square of pixels), with a depth  $z$  corresponding to different sources of information (e.g., RGB colors in images) and uses the patterns observed as it slides across the input image to set weights. The main advantages of the convolutional layers are that these layers reduce the number of parameters via their weight-sharing mechanism and gradually build up spatial and configurational invariance.<sup>245</sup> Pooling

layers essentially implement subsampling to reduce the impact of noise and the number of learned parameters. In addition, pooling layers add a certain level of resilience to small shifts in the placement and orientation of input features. A convolutional layer followed by a pooling layer can form a convolutional module, and each module of the CNN network learns to identify features while preserving their spatial relationships as shown in Figure 11. These properties lead to a major advantage of CNNs over standard ANNs, namely, that they are translation-invariant, i.e., they can recognize the same feature in different areas of the input field.

Generally, a CNN is composed of a stack of convolutional modules to achieve feature extraction, followed by one or more fully connected layer(s) for prediction and loss-minimization. It should be noted here that the convolution operation is linear, since the output of each neuron  $\phi$  in the feature map is simply the result of multiplying the input value  $x_i$  by the weights  $w_i$  of a given filter and adding them:

$$\phi = \sum_i x_i w_i$$

Therefore, in most cases, the output of a convolution layer will be passed through some form of nonlinear activation function (such as rectified linear unit, ReLU) to allow the network to handle more intricate relationships. As CNNs grow ever deeper, a new problem arises: information about the input or gradient can vanish and “wash out” when it reaches the end of the network. Several different approaches have been proposed to address this and related problems, such as creating short paths from early layers to later layers<sup>227,246,247</sup> or connecting each layer to every other layer in a feed-forward fashion to ensure maximum information flow between layers in the network.<sup>248</sup> Similar to other neural network architectures, hyperparameter tuning has a great impact on the performance of CNN models, and the number of possible choices makes the design space of these architectures large, rendering an exhaustive manual search infeasible. In this case, various approaches have been developed to provide reasonable initial hyperparameter values and to rationalize the tuning process, examples of which include grid search, random search,<sup>249</sup> Bayesian optimization,<sup>250,251</sup> and evolutionary methods.<sup>252</sup>

CNNs were originally developed for two-dimensional image recognition. Some of the approaches utilized in typical CNN networks, such as the pooling algorithm, are adopted to reduce the dimensionality of the representation and to allow for position shift, which can lead to a loss of information and hence a poor performance in drug discovery-related studies. An advanced deep learning architecture, the Capsule Network,<sup>253</sup> allows the modeling of hierarchical relationships of the network’s internal knowledge representation and could have considerable potential for drug discovery-related research.

**2.1.2.6.6. Generative Adversarial Network and Its Variants.** In a generative adversarial network (GAN),<sup>254</sup> two independent competing networks (Figure 12) are trained simultaneously: one is the Generator ( $G$ ), which takes an input  $\mathbf{z}$  from probability distribution  $p(\mathbf{z})$  and generates data  $G(\mathbf{z})$ , and the other is the Discriminator ( $D$ ), which receives as input either output from the generator  $G(\mathbf{z})$  or examples from the training data  $\mathbf{x}$  and acts as a binary classifier to predict whether the input is real or generated. Over the course of simultaneous training, the weights and biases in the discriminator and generator are updated through backpropagation. As the

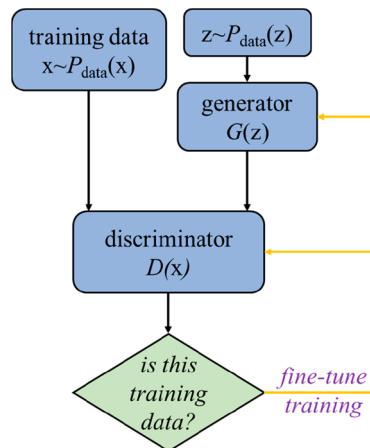


Figure 12. Architecture of generative adversarial networks (GANs).

discriminator tries to learn from its mistakes and improves its ability to discriminate between real and generated inputs, the gradient information is backpropagated to the generator to tune its weights and biases, allowing it to generate new samples which are closer to the real samples, in an attempt to fool the discriminator. This process repeats until some arbitrary criterion is met. Mathematically, the training process can be expressed by

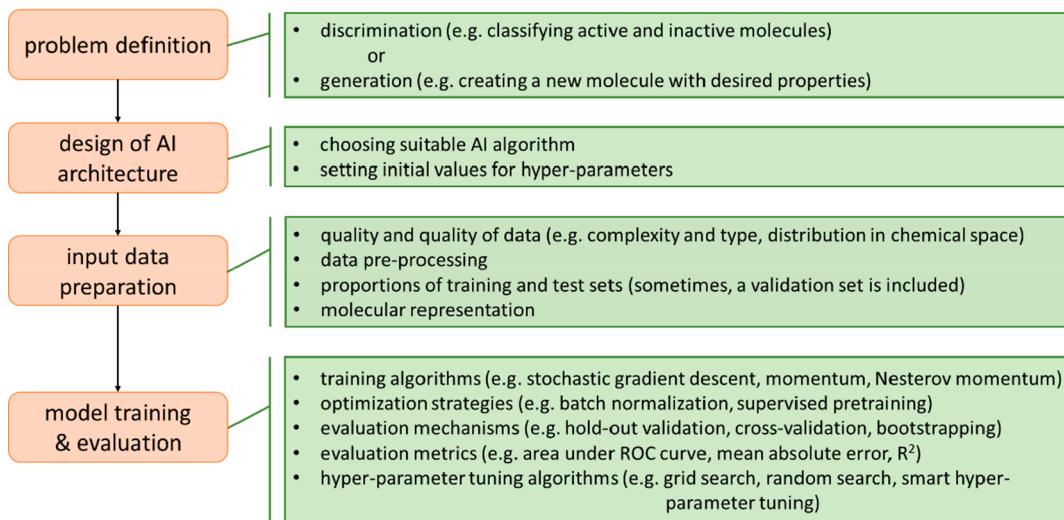
$$\min_{\theta_2} \max_{\theta_1} \{ \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \log D_{\theta_1}(\mathbf{x}) + \mathbb{E}_{\mathbf{z}} \log(1 - D_{\theta_1}(G_{\theta_2}(\mathbf{z}))) \} \quad (10)$$

where  $p_{\text{data}}(\mathbf{x})$  is the underlying distribution of the training data. The training process continues until the generated samples are (or are nearly) indistinguishable from real data.

GANs suffer from several difficulties, e.g., mode collapse and lack of informative convergence metrics, and so on.<sup>255</sup> Arjovsky proposed a new algorithm named the Wasserstein GAN (WGAN),<sup>256</sup> in which the Earth-Mover distance is used to replace the original Jenson-Shannon divergence, such that the gradient will not vanish when the generator reaches an area where the discriminator has a high degree of confidence. Additional problem-specific metrics, such as the perceptual metric used in a popular paper on the reconstruction of down-sampled images,<sup>257</sup> can help tailor these and other methods to improve their output in a more useful fashion.

**2.1.2.6.7. Attention and Attention-Based Architectures.** Rather than being a network architecture as such, attention is a means of improving the performance of other models, with special attention to the generative RNN and encoder-decoder architectures, and allowing them to generate more useful output in cases where there are substantial long-range dependencies in the input or output sentences. English-to-German translation is often given as an example of such a problem, given that German is a subject-object-verb order language. This means that decoders can struggle to produce sensible input, as they are not sure of the relationship between subject and object.

Bahdanau et al.<sup>258</sup> developed an approach to get around a “bottleneck” introduced by encoder compression, forcing a fixed number of dimensions onto the latent space representation of input sentences and avoiding the direction-dependence of earlier methods,<sup>259</sup> significantly increasing performance on translation tasks. In essence, the attention model allows the decoder in a bidirectional-LSTM encoder/LSTM decoder



**Figure 13.** General workflow of constructing an AI predictor.

model to focus on sections of the input when producing an input, thus “paying attention”, rather than using only information in the context vector. Luong et al.<sup>260</sup> extended this work, adding new kinds of attention models, and the distinction between global and local attention.

## 2.2. Artificial Intelligence Modeling in Drug Discovery

**2.2.1. Constructing a Model.** Applying a particular AI algorithm in drug discovery is a sequential process that requires the proper definition of the problem domain.<sup>261</sup> This process typically includes problem definition, data preparation, design of the AI architecture, model training and evaluation, and understanding and explaining the results (Figure 13). More specifically, one should be clear about the problem at hand before any specific architectural decisions are made (step 1 in Figure 13), since the choice of the machine learning method should be appropriate for the problem under study. First, one needs to know whether this particular problem belongs to the domain of discriminative or generative tasks. With the task of AI modeling in mind, the next step is to design an appropriate model architecture (step 2 in Figure 13). This step includes the choice of a suitable algorithm and setting sensible initial values for hyperparameters. For discriminative tasks, SVM, RF, and ANNs are the most frequently used algorithms to date. Various models are available for generative tasks, such as DBM, DBN, GAN, VAE, and AAE. In general, ANNs are the go-to part of the field for these tasks, given their ability to generalize and hypothetical ability to approximate any input-output relationship function. Hyperparameters vary with different AI algorithms. For example, in SVM algorithms, this refers to the choice of kernel, the kernel parameters (such as  $\gamma$  in the case of RBF), and the error penalty  $C$ . The architectural parameters for a neural network include, but are not limited to, the choice of the number of neurons and layers (and their type), learning rate and decay, regularization parameters, and the presence of connections between neurons or adjacent layers.

After a provisional architecture is determined, it is time to prepare the data set (step 3 in Figure 13). The representativeness, quality, and quantity of initial data have a crucial impact on the quality of an AI model. Once the initial architecture and the data sets have been established, one may proceed to model training and evaluation (step 4 in Figure 13). The training step

aims to search a set of parameters with the objective of reducing/minimizing the prediction error. The final AI model should have the ability to express the underlying relationship between the molecular representations and practitioners’ own specific purposes. If this is not the case then examining specific examples can help guide the practitioner in developing their model “ecosystem” to accomplish the goal.

**2.2.2. Input Data Preparation.** Newcomers to the application of AI methods in drug discovery projects tend to make improvements on the overall performance of the model by focusing on the deployment of the latest AI approaches. However, it is often more beneficial to focus in the first instance on the training data, as it underpins all further progress. More high-quality data usually leads to a better generalization performance, regardless of model chosen.<sup>262</sup> However, data preparation is a labor-intensive and challenging task. One needs to understand the origin and meaning of the training data, for example, the types and complexity of entities represented, the quantity of data, and more domain-specifically, their distribution in chemical space, for instance. Overall, the question is how well we have populated the space of possible inputs that we might want to make predictions for. If the need for more data is apparent, one must decide on a preprocessing strategy, whether unlabeled data might suffice, what sort of representation would be most useful for encoding the entities represented, etc. Importantly, no single rule is universally applicable. Here, we attempt to provide some guidance on these issues.

**2.2.2.1. Data Types in Medicinal Chemistry.** Table 2 presents data types used as inputs ( $X$ ) or outputs ( $Y$ ) for building AI models in drug discovery.<sup>263</sup> The most frequently used input data type in drug discovery is a fixed-length input vector (e.g., molecular descriptors, fingerprints).<sup>264,265</sup> However, there are two major limitations inherent to this type of representation. Such vectors tend to be rather large to encode all possible substructures without collisions (overlap), resulting in models that have many learnable parameters and that attempt to learn from relatively sparse inputs. For example, a fingerprint vector of size 43,000 was used by Unterthiner et al.<sup>266</sup> Similarity assessment in such a high-dimensional chemical space is error-prone.<sup>128</sup> To partially alleviate this problem, various types of graph fingerprints<sup>267,268</sup> have been

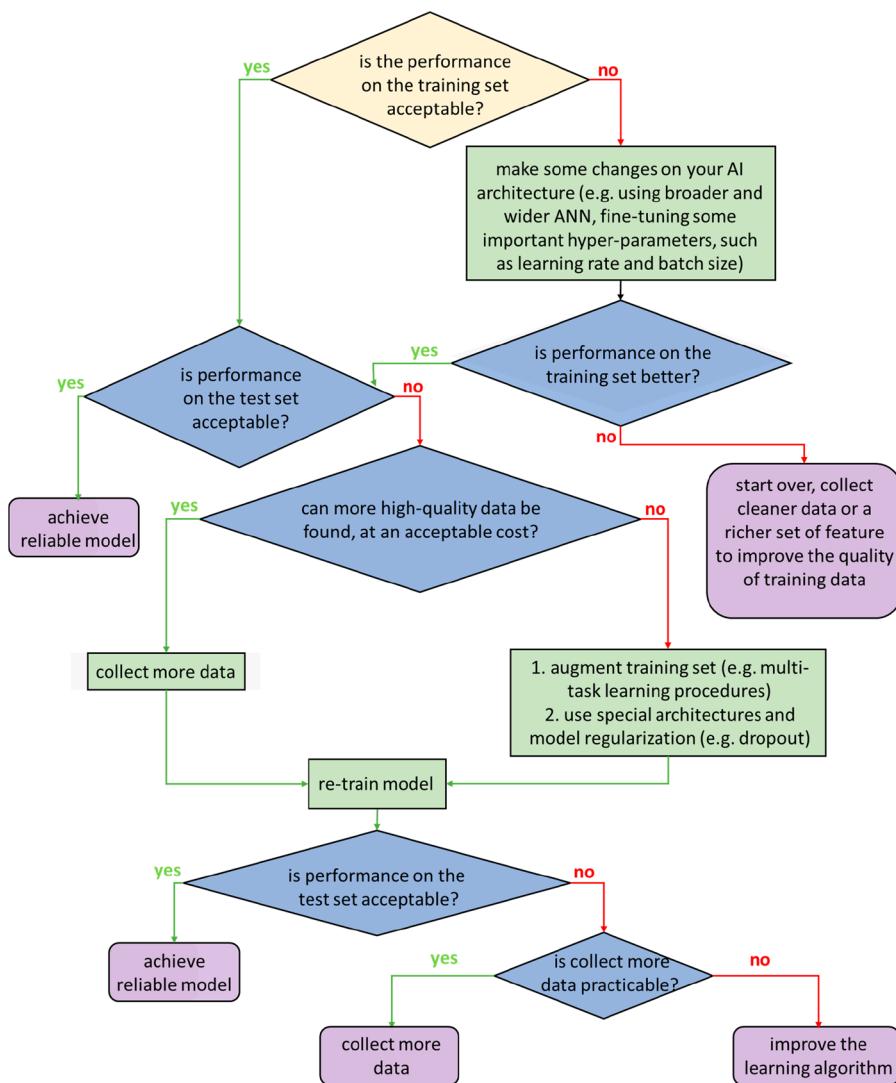
**Table 2. Different Types of Data Used in Drug Discovery:  
Input X and Output Y**

data type	
input X	sequence (e.g., primary structures of biomacromolecules, biomedical texts for extracting drug–drug interactions; SMILES strings, time-series data) fixed-size vector (e.g., bitstrings, vector of real-valued numbers) molecular structure graph
output Y	binary values (binary classification tasks) integer values (multiclass classification tasks) real-valued numbers (regression tasks) fixed-size vector (e.g., a generated fingerprint vector) single data column corresponds to single-task learning; multiple data columns correspond to multitask learning sequential data (e.g., SMILES string, amino acid sequence)

proposed, which are calculated with a differentiable neural network whose inputs are molecular structure graphs. The other limitation is caused by the difficulty in establishing bijection (one-to-one correspondence) between input vectors and molecular structures. Input vectors can be easily generated from a molecular structure, but reverse structure reconstruction from vectors is an extremely difficult task, especially as a

single fingerprint representation likely corresponds to multiple possible chemical structures. One way to avoid this limitation is to use AI deep generative models (for more details, see section 2.1.2.6.6) in combination with SMILES strings as molecular representations.<sup>269</sup> Such combinations have been widely investigated in recent years for the generation of novel compounds with desired properties (for more application details, see section 5). In this case, the output data is either a SMILES string or a fingerprint vector. The output data in most other AI-assisted drug discovery projects are numerical value(s), with binary values corresponding to binary classification, integer values for the multiclass classification (or clustering), and real-valued numbers involved in regression tasks, often with experimental biochemical data.

**2.2.2.2. Quality and Quantity of Initial Data.** One should pay special attention to the representativeness and the cardinality of the initial data. If the data set size is not representative of the underlying learning task, the training process tends to produce overfitting and suffer from poorly conditioned average error, both of which increase the likelihood of a given model being incapable of true generalization to new data. But how does one judge whether



**Figure 14.** An algorithm to guide data-collection and model-improvement strategies.

to collect more data? Here are several suggestions for interested readers (Figure 14).

One should first determine whether the training set performance is acceptable. If the training set performance is unacceptable, even with model regularization (such as dropout<sup>226</sup>) disabled, then it is likely that the architecture chosen could benefit from some tuning or reconstruction. In general, it should be possible to learn the training data with even a poorly optimized architecture, so it is best to focus efforts here rather than immediately going in search of more data. If, after tuning or trying another architecture (where appropriate), one still does not see improvement, it might be time to consider the quality of the training data, in terms of noise and errors, and check if the training distribution is balanced and the input-output correspondence is logical.<sup>270–272</sup> A further recommended test is to retrain the model with scrambled Y-data as a sanity check, where poor performance of the trained model on the scrambled data is desirable to demonstrate that the model is learning meaningful relationships in the unscrambled case.<sup>273</sup>

If the training set performance is acceptable, then one should check the performance on a test set. If the gap between training- and test-set performance is unacceptably high then collecting more data will be one of the most effective solutions. Key considerations include the relative cost of collecting more data versus improving test set performance in other ways and the amount of data that is expected to be required to significantly improve the test set performance (*n.b.*, use a statistical test to determine the significance). It can be a useful exercise to feature-stratify the input data and retrain the model with different proportions of each strata drawn uniformly, as well as randomly (or with a specific bias), to obtain an idea of the generalizability and stability of the model. It is almost always preferable to collect more high-quality raw data at a relatively affordable cost. A more typical scenario is that such approaches are costly or infeasible, which is often the case in drug discovery with biological data. The training set can be augmented by the generation of artificial data points<sup>274–276</sup> or by the use of multitask or transfer learning procedures to enrich the model with insights about related data.<sup>277</sup> The model can be improved through hyper-parameter optimization, the incorporation of regularization (especially if the model is overfitting), or through adoption of more specialized neural network architectures, depending on the task at hand. If the performance on the test set is still much worse than on the training set, even after adjusting the regularization hyperparameters, and it is impracticable to collect more training data, the easiest way to reduce the generalization error may be to improve on the learning algorithm itself. For example, the GAN algorithm makes it possible to effectively generate artificial molecule structures with desired properties from limited labeled data and abundant unlabeled data (for more details about this algorithm, see section 2.1.2.6.6).

It is worth emphasizing again that, the initial training data should be of not only a sufficient quantity but also a sufficiently high quality. However, the current situation in the drug discovery field is that our data quality is not fantastic nor is the quantity that large. Taking binding affinities as an example, maybe there are 2000 “reliable” values available in the literature (though databases give ~10,000 with a mix of IC<sub>50</sub>, K<sub>d</sub>, etc.). For the sake of argument let us say we can use all 10,000 values; this amount of data is infinitesimal when compared to the vast amounts of information available to build

vision models or models that relate what you purchased to what else you might want to purchase. Furthermore, databases generated with high-throughput screening (HTS) data are often imbalanced with respect to active and inactive compounds,<sup>278–280</sup> although specific data sampling techniques (e.g., bagging techniques) can be used to balance the distribution of activities for modeling.

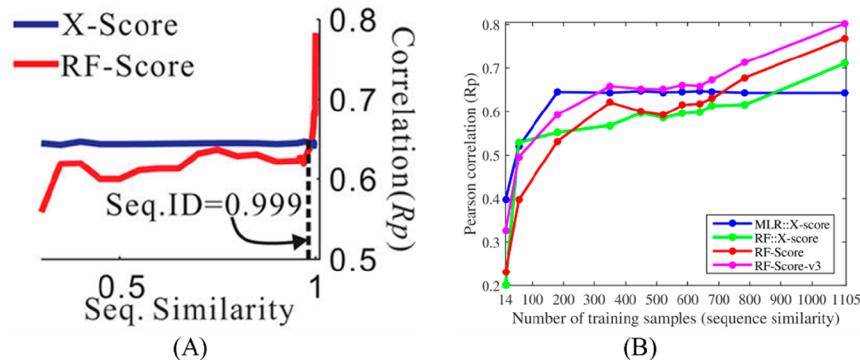
One shoud keep in mind that unlike the data in, for instance, computer vision where the data is reliable and the underlying problem is fully described by the data, after all a cat is a cat and not a dog, the data sets used for machine learning in drug discovery do not necessarily permit a solution to be found by the learning system. We have to concede that our understanding of the domain and consequently the respective data sets are incomplete and they contain error. Furthermore, it is well-known that most drugs have multiple biological targets and activities, and the individual genetic profile of patients defines their relative importance. This means, that in certain areas of drug design we are confronted with inherently ill-posed problems because of unknown contributing factors and many-to-many nonlinear relationships.<sup>262</sup>

The confidence that can be placed in the labeling of compounds as inactive against a given target also depends on its source. For example, some widely used resources, such as the directory of useful decoys (DUD),<sup>281</sup> assign compounds an “inactive” label without experimental validation. This problem can be partially addressed by the use of semisupervised methods to incorporate information on active and unlabeled compounds during training, by bootstrapping,<sup>282</sup> or by the use of noise-adapted neural nets that estimate the likelihood of an existing label being incorrect and use this to weight the data accordingly.<sup>283</sup>

### 3. APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN STRUCTURE-BASED VIRTUAL SCREENING

In the pursuit of reducing the cost of HTS while increasing its efficiency and predictability, virtual screening has been employed successfully to optimize the hit finding workflow.<sup>284,285</sup> With their strong learning and generalization capabilities, machine learning models implementing aspects of AI methods have been successfully applied in several aspects of the virtual screening pipeline. Virtual screening can be divided into two broad categories: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS), with the former corresponding to situations wherein structural information from ligand-target binding is utilized and the latter to situations with its absence. Given the breadth of the application of AI methods in virtual screening, we will begin with a discussion of discoveries in SBVS driven by such approaches and treat AI-assisted LBVS in detail in the next section (section 4).

SBVS is intended to search and rank accessible chemical space for potential ligands on the basis of a three-dimensional structural model of the intended target macromolecule (usually a protein, or RNA structure). Molecular docking techniques have been popular in SBVS since the early 1980s.<sup>286</sup> Most molecular docking studies have been conducted with a fully flexible ligand and a constrained target because of the relatively high computational overhead of this approach, although modern hardware has rendered it practicable at a large scale.<sup>287</sup> The first step of docking is placement, in which a molecule from a preprepared library is virtually docked into a target’s binding site based on steric and physicochemical



**Figure 15.** Test set performance of scoring functions trained at different sequence similarity levels. (A) A decrease is observed when the sequence similarity between the training and test data decreases from 1 to 0.991. (B) Nested training data sets at different protein sequence similarity cutoffs and a test set with 195 diverse protein–ligand complexes were used to evaluate the performances of different scoring functions. Panel A is reproduced from ref 318. Copyright 2017 American Chemical Society. Panel B is reproduced from ref 319. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2018.

properties. In the second step, the generated binding pose is evaluated by predicting the energetic interactions between the target and ligand as a proxy for their affinity, using a mathematical model (scoring function). The means by which a pose should be scored is the subject of much debate. To date, these methods have usually utilized force-field or empirical measures, which use energy models or observed features to evaluate a pose. More recent efforts have focused on data-mining, with knowledge-based<sup>288</sup> and machine-learning driven scoring functions. Scoring is usually followed by a postprocessing step, wherein candidate solutions are ranked based on calculated scores and other criteria, and often, only a fraction of top-ranked molecules will be selected for subsequent experimental assays or more expensive computational analysis. Relatively robust and accurate sampling algorithms for pose generation have been developed and are currently available in various docking softwares, such as matching algorithms,<sup>289</sup> which map a ligand into an active site of a target based on molecular shape; incremental construction algorithms,<sup>290</sup> which put a ligand into an active site in a fragmental and incremental fashion; Monte Carlo (MC) algorithms,<sup>291</sup> which generate poses by randomly changing a ligand conformation; and molecular dynamics (MD) algorithms,<sup>292</sup> which move each atom in the ligand and target in the field of atoms.

The utility of the scoring functions employed continues to be a major limiting factor in the reliability of docking approaches.<sup>293</sup> Despite decades of research and the creation of a multitude of scoring functions, there is no single scoring function which consistently outperforms others in all cases, and predicting binding affinities with high accuracy and precision remains one of the most important open problems in pose scoring.<sup>294</sup>

Many scoring functions face well-known limitations related to insufficient consideration of the conformational entropy (such as protein flexibility) and energy contribution from solvents.<sup>295</sup> Furthermore, force-field-based scoring relies on an often insufficient predetermined functional form of the relationship between the binding affinity and the variables that describe a protein–ligand complex.<sup>296</sup> Although a set of weighted parameters can be tuned to customize a scoring function to a specific target, many are stated in such a fashion that the form of the regression is invariable. This rigidity of the model leads to unsatisfactory prediction performance for complexes which do not conform to the modeling

assumptions. Therefore, novel scoring functions based on flexible nonlinear models need to be developed for improvement of the prediction accuracy.<sup>297</sup>

In recent years, AI algorithms have been introduced to SBVS by means of the construction of nonparametric scoring functions. The relationship between the feature vectors and their contributions to protein–ligand binding free energy is implicitly learned from existing experimental data in a data-driven manner, which should allow for the extraction of useful nonlinear relationships to obtain generalizing scoring functions.<sup>298,296,299</sup> AI-based nonpredetermined scoring functions such as the SVM-based ID-Score,<sup>300</sup> RF-based RF-Score<sup>296</sup> and ANN-based NNScore<sup>301</sup> have been developed to identify active ligands with improved accuracy. The extent to which these AI-based nonpredetermined scoring functions outperform classical approaches in binding affinity prediction has been highlighted in several reviews.<sup>298,302–304</sup> To avoid any confusion, in what follows, we will use the term “AI-based nonpredetermined scoring functions” when referring to data-driven, AI-based scoring functions.

### 3.1. Constructing Artificial Intelligence-Based Nonpredetermined Scoring Functions

AI-based nonpredetermined scoring functions are the subject of active research and have been validated through their application in predicting ligand-binding affinity for several biological systems. Open questions include how to best construct such models,<sup>296</sup> whether these models could benefit from novel chemical representations,<sup>305</sup> and how these systems behave with variable feature representations.<sup>306</sup> In addition to obtaining the answers to these questions, it is necessary to properly validate these models<sup>307,308</sup> to avoid promoting novelty at the expense of accuracy.<sup>309</sup> Real-world performance has to be assessed.<sup>310</sup> An important step toward better scoring functions would be the introduction of innovative conformational sampling to reduce the risk of false ligand pose prediction,<sup>311,312</sup> along with the incorporation of such methods in user-friendly docking software.<sup>311,313,314</sup>

Many applications of AI techniques for the development of scoring functions have been focused on in five major techniques, namely naive Bayes, SVM, RF, feed-forward ANNs, and deep neural network approaches. An overview by Ballester and co-workers<sup>298</sup> covers the use of machine learning regression algorithms in the development of AI-based nonpredetermined scoring functions to improve protein–

ligand binding affinity prediction. To avoid repetition, we have deliberately omitted the examples which are discussed in that work<sup>298</sup> and confined ourselves to developments in this field since 2015, although a few previously covered examples remain for the sake of completeness.

Of particular note is a series of studies by Ballester and Mitchell, who developed RF-Score software.<sup>296,305,315</sup> Several other RF-based scoring functions have been developed and tested in the prediction of protein–ligand binding affinities, such as SFC<sub>score</sub><sup>RF</sup>,<sup>316</sup> B2BScore,<sup>317</sup> and RF-IChem.<sup>309</sup> Among different RF-based scoring functions, RF-Score shows sustained performance and has been incorporated into istar, a web platform for large-scale protein–ligand docking.<sup>314</sup> The istar platform combined with the retained RF-Score can achieve a Pearson's correlation coefficient,  $R_p$ , of as high as 0.86 between predicted and experimentally measured binding affinities for targets in the PDBbind v2012 refined set ( $N = 2897$  complexes). Li et al.<sup>318</sup> recently undertook closer examination of the scoring power of RF-Score-v1. They found that when the sequence similarity (defined as the number of aligned identical residues divided by the length of the test protein) between training and test data decreases from 1 to 0.99, the performance of RF-Score-v1 significantly drops from  $R_p = 0.78$  to  $R_p = 0.68$  (Figure 15A). The authors concluded that the outstanding performance of RF-Score-v1 over X-Score can be attributed to the high similarity between the training set and the test set (TM-score >0.98). However, in a follow-up article, Ballester et al.<sup>319</sup> revisited this question and found that RF-Score-v3 performed better than X-Score (which has the best test set performance out of 16 classical scoring functions<sup>320</sup>) in retrospect, even when 68% of the most-similar proteins were removed from the training set (Figure 15B). The superior performance of AI-based nonpredetermined scoring functions is partly due to model training with compounds that differ from the test set.

Various SVM-based models have been developed, such as SVR-Score<sup>321</sup> and ID-Score.<sup>300</sup> For more details on the prediction performance of those SVM-based scoring functions, we refer the reader to the review from Ain et al.<sup>298</sup> It should be mentioned that RF-based scoring functions have tended to outperform SVM-based scoring functions in retrospective benchmarks. The efficiency of RF methods over SVM methods in predicting protein–ligand binding affinity may be explained by the ensemble character of RF. Using a decision tree as base learner allows the algorithm to incorporate its naturally high flexibility and variance. The high variance can reduce the correlation between trees, thereby improving the accuracy of the whole ensemble model. Feature sampling has proven to be effective in models constructed based on tree-based ensemble algorithms such as RFs.<sup>64</sup> Both RF and boosted regression tree scoring functions have demonstrated substantial improvement in performance over a panel of 16 classic scoring functions,<sup>312</sup> with improved accuracy on the 2007 PDBbind benchmark.<sup>322</sup>

Various ANN-based scoring functions have been developed for the rescoring of docking poses, such as NNScore<sup>301,323</sup> and CScore.<sup>324</sup> Although other machine learning methods (including RFs, SVMs, and shallow neural networks) have previously been used to predict protein–ligand affinity,<sup>298</sup> a limitation of many of these approaches is the need to represent molecules with fixed-length vector features. Recent advances in deep learning algorithms, particularly CNNs, have made it possible to predict protein–ligand binding affinity by the automatic extraction of features directly from their two- or

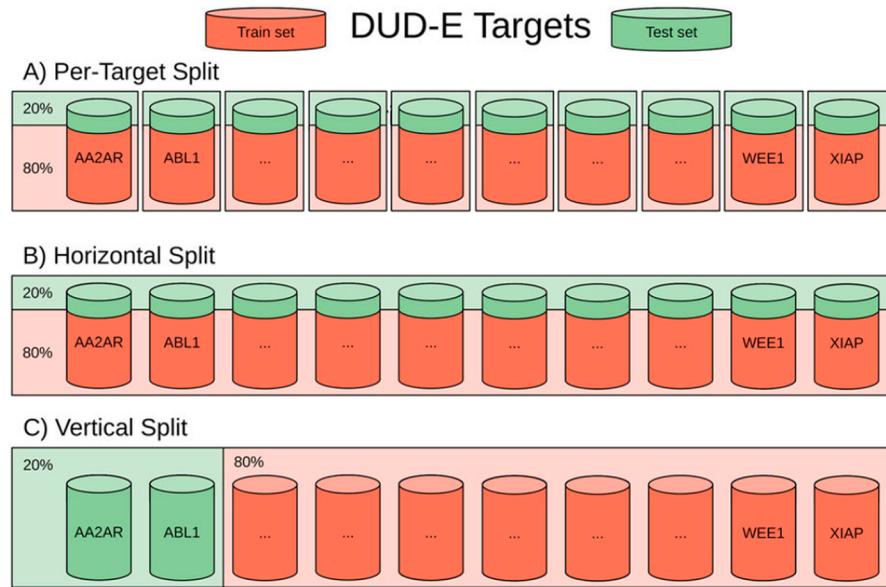
three-dimensional structure. Examples include the multi-channel topological neural network (TopologyNet) by Cang and Wei,<sup>325</sup> in which a topological strategy is used to represent the three-dimensional biomolecular geometry of one-dimensional topological invariants in a reduced-dimensionality formulation while still maintaining important biological information, and the CNN model from Ragoza et al.,<sup>326</sup> which inputs a 3D grid of each protein–ligand complex, wherein every grid point stores atom densities. AI algorithms can also be integrated into other techniques. For example, van Dijk et al.<sup>327</sup> described the use of a workflow based on iterative linear interaction energy (iLIE) for binding-affinity prediction for cytochrome P450 aromatase (CYP19A1) inhibitors. Cang et al.<sup>328</sup> proposed a strategy that integrates machine learning with homology estimation for ligand affinity prediction.

Overall, the reliable quantitative prediction of ligand affinity or potency remains a challenge. AI algorithms, specifically deep learning algorithms, have resiliency with respect to noise and extend naturally to high-dimensional data, such as protein–ligand complexes. AI models implementing nonpredetermined scoring functions have made substantial improvement over classical approaches in prediction benchmarks.<sup>298,310,329</sup> Importantly, an additive (linear) scoring function can easily be improved by replacing the respective linear regression method with nonlinear machine learning methods, e.g., with RF<sup>315</sup> or support vector regression,<sup>299</sup> or with more complex AI models such as DNNs and CNNs.<sup>326</sup> There is no single best AI-based, nonpredetermined scoring function that can outperform all others in every respect. Indeed, combining scoring functions can improve the success rate of virtual-screening projects.

### 3.2. Structure-Based Virtual Screening with Artificial Intelligence Approaches

SVM-based methods have been prominently used to develop a target-specific SVM-based scoring model (SVM-SP)<sup>330</sup> for the identification of small-molecule inhibitors among the protein–ligand complexes generated by automated ligand docking.<sup>331</sup> The inclusion of the statistical pairwise potentials of docked protein–ligand pairs has led to a general scoring function (SVMGen).<sup>332</sup> MIEC-SVM<sup>333</sup> is an SBVS approach that uses an SVM model to discriminate between active and inactive molecules. MIEC-SVM has been evaluated in several retrospective virtual screening studies and has demonstrated good performance.<sup>334,335</sup> Recently, Sun et al.<sup>336</sup> have reported a follow-up study on how to best generate an MIEC-SVM model. They found that by combining molecular docking, ensemble minimization, and molecular mechanics/generalized Born surface area free energy decomposition and hyperparameter optimization for SVM, the performance of the optimized MIEC-SVM model can achieve hit rates of up to 14% in retrospective benchmarks.

The usefulness of SVM methods for postdocking has also been thoroughly investigated. For example, a combinatorial ensemble docking scheme has been proposed by Leong et al.<sup>337</sup> to select binding poses and predict the binding affinity of NMDAR GluN1-ligands. Yan et al.<sup>338</sup> introduced a postdocking classifier (PLEIC-SVM) in which the protein–ligand empirical interaction components (PLEIC) fingerprint is used for representing the interactions found in each complex. Rodríguez-Pérez et al.<sup>339</sup> investigated the influence of the training set composition and size on the SVM-based prediction of active compounds, finding that approximately 50 active



**Figure 16.** Overview of the quality assessment strategy adopted by Wójcikowski et al.<sup>342</sup> Data from the DUD-E database were divided into three separate approaches to assess the reliability and generalizability of the RF-score-VS approach when trained on different subsets. (A) Per-target splitting, wherein each training and test set contains data from a single protein target; (B) horizontal splitting, wherein the training and test sets contain data from all targets; and (C) vertical splitting, wherein no shared targets exist between the training and test data. Each barrel represents all the target-ligand complexes (actives and decoys) for each target. Reproduced with permission from ref 342. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

compounds per target are necessary to make useful predictions and that substantial inactive data has considerable benefit in improving performance.

SVM-based integrated workflows have been designed, capitalizing on known strengths and limitations of both structure- and ligand-based virtual screening. For example, Xie et al.<sup>340</sup> combined SVM- and docking-based methods (SB/DB-VS approach) to identify novel c-Met tyrosine kinase inhibitors from a pool of 18 million compounds. Meslamani et al.<sup>341</sup> designed an automated workflow, PROFILER, to predict the most likely targets of bioactive compounds. PROFILER integrates four ligand-based approaches (SVM binary classification, support vector regression affinity prediction, nearest neighbors affinity interpolation, and three-dimensional similarity search) and two structure-based approaches (docking and protein–ligand-based pharmacophore searching).

RF-Score<sup>296</sup> has been successfully applied in several SBVS applications. The performance of RF-Score in virtual screening was benchmarked across a series of targets. A ready-to-use scoring function, RF-Score-VS, was trained on the full directory of useful decoys-enhanced (DUD-E) data sets (15,426 active and 893,897 inactive molecules docked to a set of 102 targets).<sup>342</sup> Three docking tools, five conventional scoring functions, and three versions of RF-Score-VS were used for the building and performance assessment. Three methods for generating stratified 5-fold cross-validations were utilized for assessing the model's quality (Figure 16). The RF-Score-VS top 1% accomplished a 56% virtual hit rate, whereas the best conventional scoring function Vina managed 16%. These results suggest that substantial improvement in virtual screening performance can be achieved by machine learning.

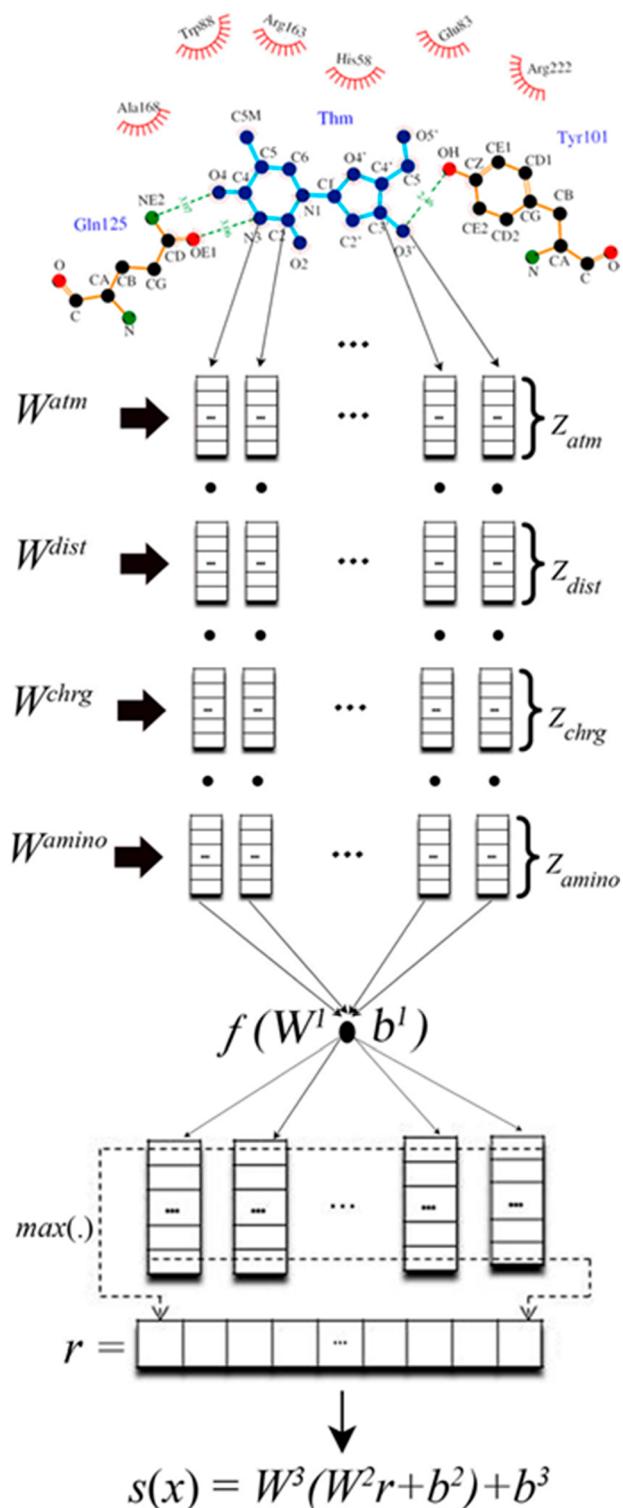
SBVS has predominantly been considered a binary classification problem. Several groups have reported that multilabel classification can improve the prediction accuracy by utilizing dependencies between target classes.<sup>343</sup> ANNs have also been used in SBVS for extracting features from

protein–ligand complexes and reranking the docked poses. While neural networks have been adopted by many groups to improve the performance of scoring functions and enhance the hit rate in virtual screening,<sup>344</sup> it is still unclear how different docking and scoring protocols affect the four different aspects of predictive performance (i.e., scoring power, ranking power, docking power, and screening power) for particular targets.<sup>345</sup>

Machine learning-based SBVS is largely dependent on the fidelity and complexity of the representations generated for the protein–ligand complexes and on the quality of the corresponding activity data. It remains common to visually examine the docked poses and manually extract potentially relevant features.<sup>302,346</sup> Deep learning methods usually extract and learn features (representations) in a data-driven way with little or no human participation.<sup>223</sup> Pereira et al.<sup>347</sup> presented such a deep learning model, DeepVS, to improve docking-based virtual screening, as illustrated in Figure 17, without manual feature engineering. In a benchmark study, DeepVS-ADV achieved the best AUC (0.81) reported until that time for virtual screening using the 40 receptors from DUD.

The application of AI in docking-based SBVS is a promising and ongoing research effort. A current trend in the development of AI-enhanced SBVS aims to improve postprocessing by rescoring the poses generated with docking programs using machine-learning models, with or without consensus scoring. As an example, one can enhance and extend Auto-Dock Vina with RF-Score-VS to achieve an improved virtual screening performance.<sup>315</sup> The combination of automated ligand docking and advanced machine learning algorithms can help reduce the number of both false-negative and false-positive predictions. Future work in this area is expected to include structural information and physicochemical properties (e.g., steric and electronic effects) of the target protein.

It should be kept in mind that many factors influence the overall performance of AI models, including data quality and quantity, the molecular representation chosen, and the model



**Figure 17.** DeepVS architecture for ligand docking. The output of a docking program (top) is fed into a CNN, and the network extracts features from basic structural data such as distances, atom types, and atomic partial charges that are suited to distinguish between actives and decoys. The concept of atom and amino acid embeddings is introduced into the network. A distributed vector representation of the protein–ligand complex is generated through representing the compound as a group of local atom contexts and by further processing and summarizing the information in a convolutional layer. Reproduced from ref 347. Copyright 2016 American Chemical Society.

architecture, to name a few. All of these factors need to be well chosen to allow for the design of useful scoring functions for SBVS. By data augmentation, here referring to the inclusion of similar but lower-quality or lower-confidence structural or binding data in the training set, an increased scoring-function performance can be achieved, due to incremental benefits from the data distribution and volume, noise-management, and diversity.<sup>310</sup>

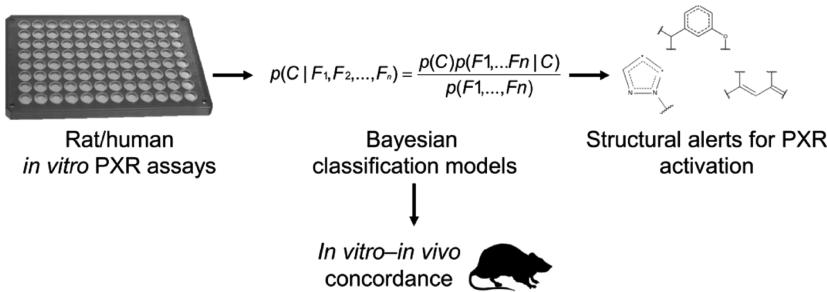
Interaction fingerprints capturing small-molecule receptor interactions are a commonly examined output of docking, as they allow the end-user to easily determine whether a given molecule has contacts with a postulated or proven key-residue; therefore, they describe the aspects of a docked pose that are generally of most interest to a modeler. A recent study demonstrated that more precise interaction fingerprints for protein–ligand complexes with complex typing schemes do not necessarily result in more accurate prediction of binding affinity.<sup>305</sup> Furthermore, the desired precision of a chosen representation is dependent on the intended domain of the AI-based nonpredetermined scoring function. For example, the predictive performance of family specific scoring functions is expected to improve when feature information such as the presence of a given metal ion is explicitly included in the model. For the time being, the suitability of AI-based nonpredetermined scoring functions is problem-dependent. The continuously increasing accessibility and abundance of curated training data implies that AI-based virtual screening approaches should, in the future, be better generalized than are the current models.

#### 4. ARTIFICIAL INTELLIGENCE APPROACHES IN LIGAND-BASED VIRTUAL SCREENING

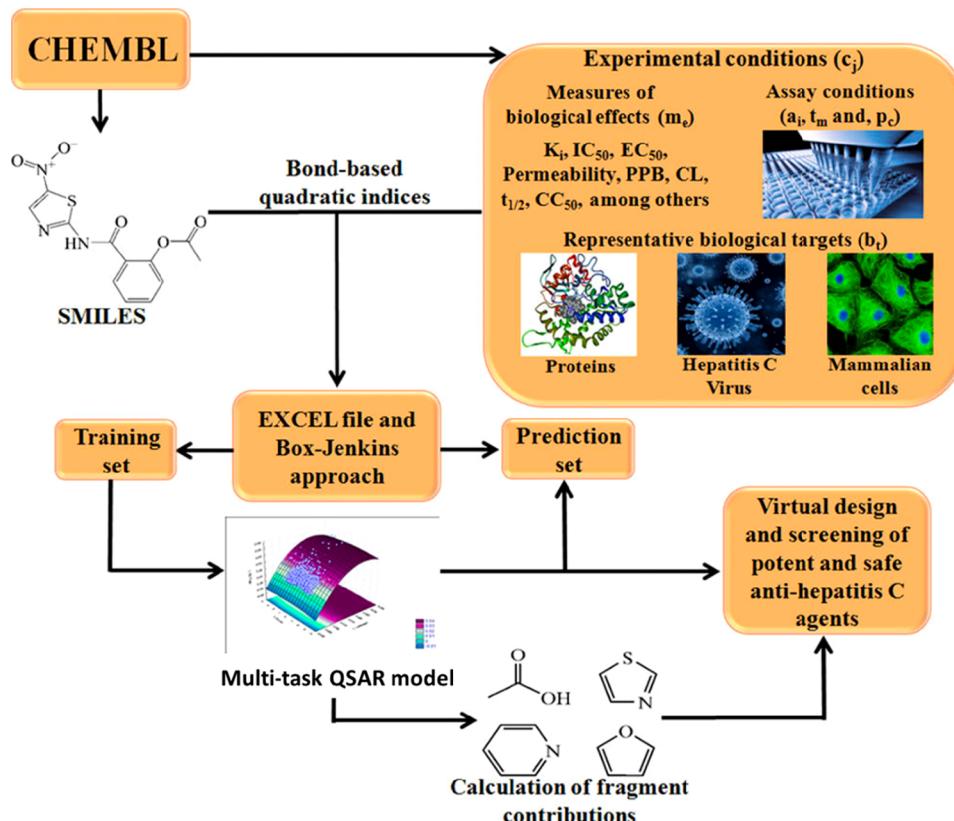
The majority of current applications of AI in virtual screening are ligand-based. LBVS is based on the hypothesis that structurally similar compounds have similar biological activities.<sup>348</sup> The general approach consists of identifying compounds known to be active against a particular target (queries), choosing the similarity method by which to compare a library of compounds to the query, and identifying candidates for further experimental assays. The AI approaches to LBVS can be broadly divided into two categories: regression models for activity prediction, and classifiers based on the some-compound-similarity metric. In this section, we survey the applicability of AI algorithms in the field of LBVS. To begin, we comprehensively review the achievements and limitations of the mainstream AI algorithms in QSAR-based LBVS (section 4.1). QSAR modeling has a long tradition in medicinal chemistry, and we refer the interested reader to some excellent reviews for key technical developments.<sup>349–352</sup> For the sake of this review, we limit the scope of discussion to the performance of, and appropriate comparisons for, QSAR models constructed from advanced AI algorithms. We then continue with compound classification-based LBVS and provide a perspective on their future in drug discovery (section 4.2).

##### 4.1. Artificial Intelligence-Based QSAR Models

The goal of QSAR modeling in drug discovery is to find a mathematical relationship that connects molecular properties (as encoded in molecular descriptors) with some quantitative compound activity metric. Implementing AI algorithms in QSAR models follows the general workflow we have described in section 2.2.1. Particularly, RF approaches have been used in



**Figure 18.** Predicting rat pregnane X receptor (PXR) and human PXR activators using ligand-based Bayesian QSAR models. Reproduced from ref 376. Copyright 2016 American Chemical Society.



**Figure 19.** Example of a multitask QSAR-model. The initial data set is retrieved from ChEMBL, with each molecule related to one or more experimental conditions. Topological descriptors known as bond-based quadratic indices are calculated. The best-performing multitask quantitative structure-biological effect relationship model is applied for fragment-based and virtual screening of new molecular entities, e.g., screening of entities exhibiting potent anti-HCV activity and desirable *in vitro* ADMET properties. Reproduced from ref 400. Copyright 2017 American Chemical Society.

QSAR-based LBVS, producing relatively reliable prediction performance on related data sets and a clear and readily interpretable contribution of features to the final prediction (cf. section 2.1.2).<sup>353</sup> Several RF-based regression LBVS models have been established, such as Profile-QSAR (pQSAR) 2.0.<sup>354</sup> This RF algorithm enables an amenable chemical and biological interpretation with regard to the molecular descriptors and structural motifs considered important by the model.<sup>355</sup>

SVM is another popular method in QSAR-based LBVS and has been used, among other applications, for the discovery of inhibitors of Cathepsin L<sup>356</sup> and factor XIIa,<sup>357</sup> as well as in the construction of multiple classification models for kallikrein 5 inhibitors.<sup>358</sup> Other studies have focused on the impacts of the

data set size and the parameter values on the LBVS performance of SVM QSAR models.<sup>359–361</sup>

Bayesian algorithms are another well-known method for constructing QSAR models (cf. section 2.1.2).<sup>362–364</sup> In general, we refer to naïve Bayes/Laplacian-modified naïve Bayes when we discuss Bayesian methods in this field. Previously, this approach has been used to identify inhibitors of general and specific kinases,<sup>365</sup> (key and/or preclinical) targets against Alzheimer's disease,<sup>366,367</sup> G-protein-coupled receptors,<sup>368</sup> *Escherichia coli* dihydrofolate reductase,<sup>369</sup>  $\gamma$  amino butyric acid type A (GABA A) ionotropic receptor,<sup>370</sup> and *Mycobacterium tuberculosis*,<sup>371,372</sup> and to discover important structural features of microsomal stability.<sup>373</sup> Bayesian models have also been widely used for predicting pregnane X receptor activators. For example, Pan et al.<sup>374</sup> described the

discovery of potential therapeutic pregnane X receptor activators, such as nifedipine, using a Bayesian QSAR classification model developed with a training set of 177 compounds. Recently, Shi et al.<sup>375</sup> developed a Bayesian classification model to spot pregnane X receptor activators by using a set of 532 structurally diverse compounds, yielding a prediction accuracy of 85% for the test set. AbdulHameed et al.<sup>376</sup> constructed ligand-based Bayesian QSAR models for both rat and human pregnane X receptor activation, which allowed them to predict human pregnane X receptor and rat pregnane X receptor activation (Figure 18).

ANNs are one of the most popular tools for QSAR research. For example, ANN models were used to select new antibacterial compounds<sup>377</sup> and identify novel antibacterial 3-hydroxypyridine-4-one antibiotics with inhibitory activity against *Staphylococcus aureus*.<sup>378</sup> This modeling concept has been applied to nonsmall-molecule compounds and used to virtually screen and rank a large library of peptide candidates for the identification of small antimicrobial peptides effective against a broad spectrum of antibiotic-resistant superbugs,<sup>379,380</sup> as well as to connect peptide physicochemical and antimicrobial properties.<sup>381</sup> Other applications across the broad field of QSAR approaches have included its use to investigate a series of HEPT derivatives as potential inhibitors of HIV-1,<sup>382</sup> predict cannabinoid receptor ligand affinity,<sup>383</sup> discover new bioactive compounds against *Mycobacterium tuberculosis* and HIV,<sup>384</sup> design aldose reductase inhibitors,<sup>385,386</sup> classify the antihepatitis C activity of thiourea derivatives,<sup>387</sup> predict the antimicrobial activity of quinolone derivatives,<sup>388</sup> discover novel M5 inhibitor chemotypes,<sup>389</sup> etc. There is a multitude of potential applications for this combination of well-established and novel methodologies.

Several groups have demonstrated the usefulness of multitask QSAR in virtually screening compounds with dissimilar biological activities.<sup>390–393</sup> Multitask QSAR models classify compounds based on not only their biological effects but also experimental information (Figure 19). Tenorio-Borroto et al.<sup>394</sup> reported the construction of a multitask QSAR model based on ANN. This model classified the data set obtained from multiplexing assays with an accuracy of 92%. Follow-up investigations have been published in which various descriptors were used to generate new multitask QSAR-ANN models with the aim of predicting the immunotoxicity of chemicals.<sup>395–397</sup> The first work devoted to the discovery of safe antibacterial drugs by this approach was reported by Speck-Planche et al.<sup>398</sup> This model achieved high prediction accuracies for both the training and prediction sets. By following the same methodology and conceptual process, several promising multitask QSAR models have been constructed.<sup>393,398–402</sup>

Several deep learning algorithms have been used for multitask QSAR.<sup>266,403</sup> For example, Dahl et al.<sup>52</sup> investigated the dependency of multitask QSAR on the size and heterogeneity of the training data, finding that adding additional complexity, via hidden layers, is not necessarily useful with limited data, because of the ratio of network parameters to training examples, when using simply feed-forward networks.<sup>404</sup> Unterthiner et al., who used multitask modeling through deep learning,<sup>266</sup> used a data set of two million data points for 1280 biological targets, adding several modifications to improve the training performance and finding that the model out-performed various popular methods, including SVM, kNN, and a similarity ensemble approach (SEA).<sup>405</sup> Pande and co-workers<sup>406</sup> focused on the influence of

the number of tasks in the multitask classification problem. Over 200 targets (mostly proteins) and 37.8 million experimental data points for 1.6 million compounds, described as extended-connectivity circular fingerprints (ECFP), were collected. The multitask network outperformed logistic regression, RF, and single-task neural networks. Although performance improvements resulting from multitask DNNs have been reported by many different groups, few studies have focused on the basis for this effect. Xu et al.<sup>407</sup> found that a task embedded in a multitask DNN can “borrow” information from other QSAR tasks during the training process.

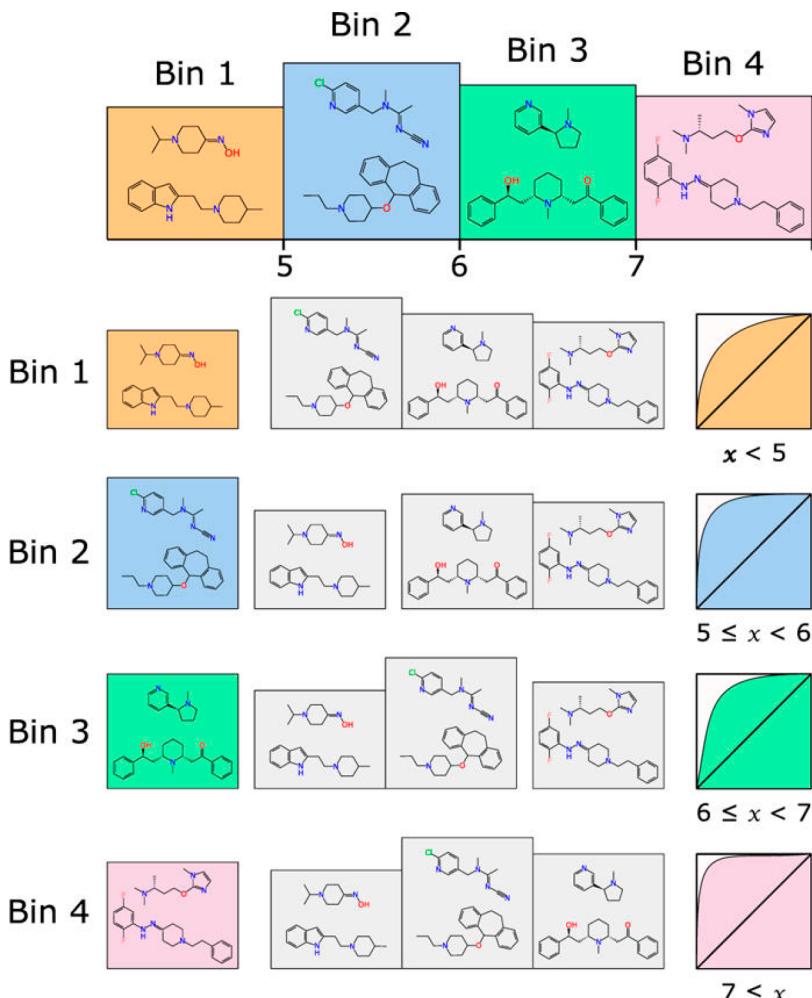
Many publications have demonstrated the improved performance of DNNs over traditional machine learning methods in constructing high-performance QSAR models.<sup>408</sup> How to optimally adjust the hyperparameters of DNN architectures remains an open question. A recent study<sup>409</sup> showed that certain hyperparameters greatly affect the performance of DNN models, including the activation function, dropout regularization, number of hidden layers, and number of neurons per hidden layer. Particularly, the dropout technique has been recommended for the training of DNN models.<sup>410</sup> Both evolutionary and Bayesian techniques have been suggested for proper parameter tuning,<sup>411</sup> to avoid limitations that might be placed on the performance of these models by heuristic approaches.

#### 4.2. Other Artificial Intelligence-Assisted Classification Strategies

In addition to the widely used QSAR approaches, in which AI methods have demonstrated state-of-the-art performance levels (for more details, see section 4.1), there are other AI-based approaches that have been employed to improve the virtual screening of chemical databases. These include clustering- and similarity-based methods, for example. In this section, we concentrate on AI-based strategies other than the combination of QSAR and AI approaches which can be found in section 4.1. For a description of the underlying algorithmic basis, we direct the reader to section 2.1.

Apart from ANN-driven LBVS, there are many instances in which different algorithms, such as the SVM method,<sup>412–416</sup> and Bayesian inference<sup>417–419</sup> have played an active role in drug classification. There are many publications relating to the successful application of Bayesian models, for example, in the discovery of novel vascular endothelial growth factor receptor-2 scaffolds<sup>420</sup> and in identifying potential farnesoid X receptor agonists,<sup>421</sup> liver X receptor  $\beta$  agonists,<sup>422</sup> ROCK II inhibitors,<sup>423</sup> cyclin-dependent kinase inhibitors,<sup>424</sup> DNA gyrase inhibitors,<sup>425</sup> breast cancer resistance protein (BCRP) inhibitors,<sup>426,427</sup> dipeptidyl peptidase IV (DPP-IV) inhibitors,<sup>428</sup> butyrylcholinesterase (BuChE) inhibitors,<sup>429</sup> and inhibitors of  $\beta$ -hematin formation.<sup>430</sup> The generalizability of this method to new targets is well-established.

A special category of Bayesian methods, Laplacian-modified naïve Bayes (LMBN) variants, has proven especially powerful in LBVS.<sup>417–419</sup> Ekins and co-workers<sup>363,431–433</sup> used this method to construct Laplacian-modified naïve Bayes classifier models for *Mycobacterium tuberculosis* (Mtb) and cytotoxicity end points. In 2015, Clark et al.<sup>362</sup> released an open source Laplacian-modified naïve Bayesian model for in vitro and in vivo bioactivity prediction. In this model, a threshold needs to be predetermined by the user to partition the collection of training data into two states (i.e., active vs inactive). The selection of the threshold has a significant impact on the



**Figure 20.** Visual example of the binning system proposed by Clark et al.<sup>362</sup> In this case, a training set is divided into four bins with definite boundaries for each bin. The next step is to build a Bayesian model for each bin. Assessing a molecule involves submitting it to each of the Bayesian models to predict whether it belongs in the associated bin. The most likely range for this test molecule is the one with the highest calibrated prediction value. Reproduced from ref 435. Copyright 2016 American Chemical Society.

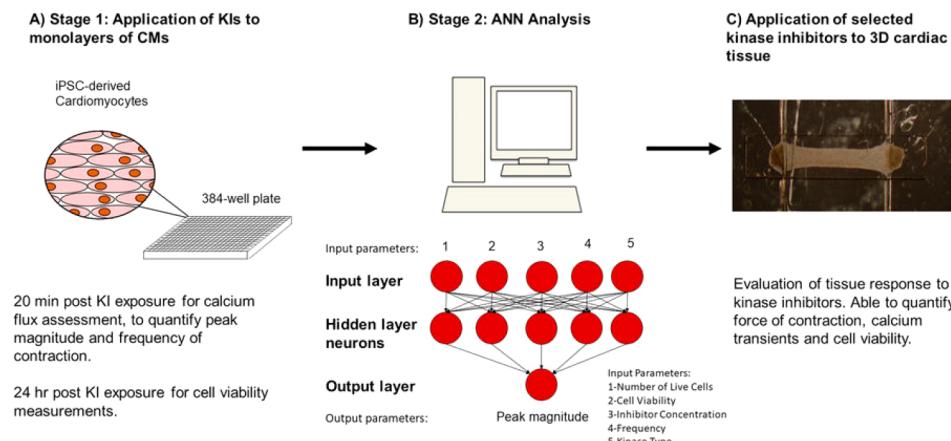
quality of the model. In a follow-up investigation,<sup>434</sup> the same group developed an automated algorithm which uses the computed receiver operating characteristic integral as an indication of partitioning efficacy for evaluating the suitability of the chosen threshold.

In addition to the requirement for an appropriate threshold, the researchers noticed that mapping the result of a Bayesian model (a pseudoprobabilistic class likelihood) to traditional QSAR or quantitative structure–property relationship (QSPR) methods, in which the outputs are in the same format as the experimental input data, is generally unsatisfactory. In some scenarios, multiclass labeling (e.g., high, medium, and low activity) is more reasonable, such as in the handling of well-determined dose–response assays. To broaden the applications of Bayesian models, a competitive approach can be taken, in which activity measurements are subdivided into multiple bins (Figure 20),<sup>435</sup> such that a composite group of Bayesian models can be created to extend the Laplacian-modified naïve Bayesian method to work with multiple states, rather than just binary cases.

The integration of Bayesian algorithms and ANNs allows the optimization of multiple control parameters simultaneously and reliably. The first publication to use Bayesian neural networks to classify central nervous system-active and -inactive

compounds was reported by Ajay et al.<sup>436</sup> as early as 1999. One group<sup>437</sup> developed a new similarity-based virtual screening method using a Bayesian inference network, which allows for estimation of the likelihood of any two compounds sharing a similar activity profile. The molecular descriptors are weighted, deprivitizing those with a less-significant relationship with the activity classes present in the training data. The results showed improved performance of the Bayesian inference network approach over that of the Tanimoto similarity method for the majority of targets in the MDL Drug Data Report. In another publication,<sup>438</sup> the same group investigated the use of Bayesian belief networks in which the root nodes represent substructural fragments and the leaf nodes represent the reference structure and the database compounds, achieving a similar performance to that of established Tanimoto-based similarity searching. This Bayesian belief network approach for activity prediction was extended to a further study,<sup>439</sup> wherein the Bayesian belief network architecture adopted was not resilient when trained on heterogeneous data and proved useful primarily as a tool for visualization and exploration of the dependencies between input variables.

The SOM architecture has a substantial track record of use in LBVS as a visualization and clustering tool (for more details



**Figure 21.** Kinase inhibitor screening workflow. First, a traditional high-throughput two-dimensional assay is conducted to screen 80 different kinase inhibitor molecules, from the GlaxoSmithKline published kinase inhibitor set (PKIs), on hiPSC- cardiomyocyte (CM) monolayers. The effects of these compounds on cell viability, cell contractility, and live cell number are chosen as the end points. ANN modeling is then used to analyze the acquired experimental data to select candidate kinase inhibitors. Finally, the selected kinase inhibitors are evaluated in the three-dimensional Biowire tissues. Reproduced with permission from ref 450. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

of this algorithm, see section 2.1), with implementations in several popular chemical softwares.<sup>440,441</sup> Many examples of their successful applications to large data sets in LBVS have been reported. For example, the Schneider group previously developed a SOM-based prediction of drug equivalence relationships (SPiDER) software,<sup>152</sup> which tessellates a reference ligand space<sup>442</sup> and predicts targets for new compounds based on assigning them to a particular winner neuron, then performing consensus scoring and statistical analysis to estimate the pseudoprobability of activity against a range of targets. Building on the success of this approach, the SOM concept was extended by the same group for many different tasks, grouping molecules with similar pharmacophoric features into clusters of functionally related compounds to generate target predictions which were then experimentally validated.<sup>443–447</sup> In addition, a similar technology, the counter-propagation ANN (CP-ANN),<sup>448</sup> was used in a publication by Merk et al.<sup>449</sup> to identify novel modulators of the farnesoid X receptor, based on a combined approach using kNN and pharmacophoric screening. This study demonstrated the power of this approach for introducing useful ordering and segmentation of chemical spaces.

An example of the combination of ANNs and HTS was recently reported by Conant et al.<sup>450</sup> They designed a workflow (Figure 21) to screen kinase inhibitors for human cardiac physiology and function. This study showed that the integration of ANNs with cell-level experiments can help to efficiently refine the set of potential molecules for expensive and time-consuming three-dimensional tissue testing.

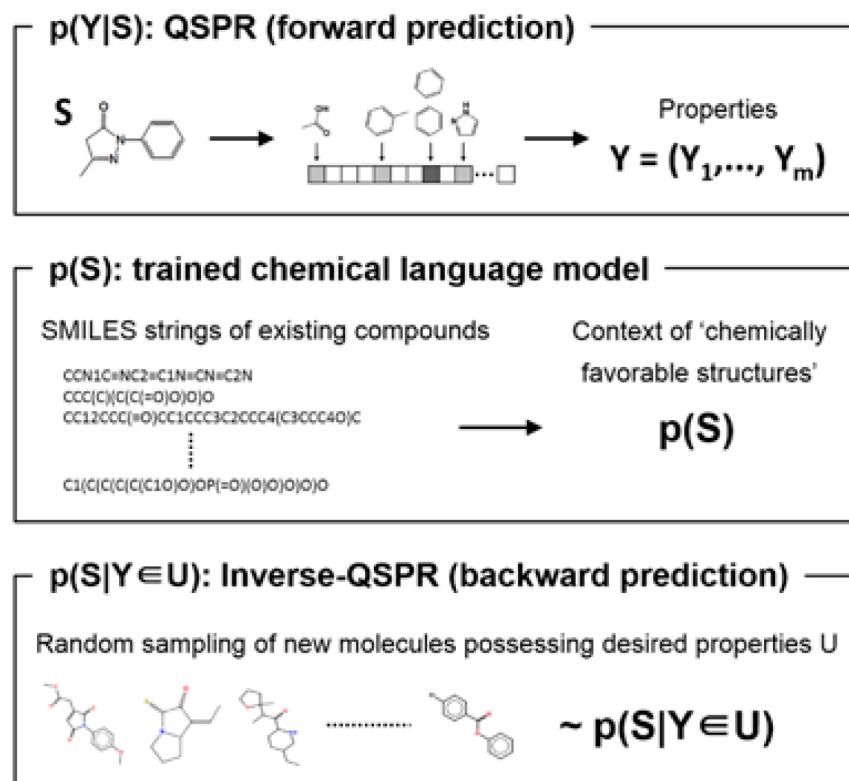
Overall, the increase in sophistication of the choice and design of AI algorithms, coupled with ever larger data sets, has contributed to their growing stature in LBVS. The comparison of different machine learning models has historically been difficult because of the lack of well-standardized assessments. Attempts to put this on a sounder footing for chemistry-related issues in AI are ongoing.<sup>451</sup> For the purpose of constructing generalized and robust AI-based predictors, it is important to provide a formal definition of the applicability domain for a developed model. Another consideration is the importance of selecting informative molecular representations for the task at hand, an undertaking which has been much improved by

recent algorithmic and computational advances. In recent years, sophisticated molecular representation methods have been developed and introduced to the field. For example, “neural fingerprint”<sup>267</sup> methods have been developed, utilizing CNNs to represent small molecules as finite-sized feature vectors. At a different level of abstraction, molecular quantum mechanical calculations have become much more routine in recent years<sup>452,453</sup> and can create reasonably high-fidelity molecular representations within a tolerable time frame. Therefore, they are attractive sources of molecular descriptors, which, in principle, can express the electronic and geometric properties of molecules and their interactions more fully than can simpler representations. In the near future, we expect that combinations of domain-specific molecular representations with AI methods will become mainstream LBVS techniques.

## 5. APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN DE NOVO DRUG DESIGN

Drug design aims to generate chemical species that meet specific criteria, including efficacy against pharmacological target(s), a reasonable safety profile, suitable chemical and biological properties, sufficient novelty to ensure intellectual property rights for commercial success, etc. With the aid of novel algorithms to carry out the design and evaluation of molecules in silico, de novo drug design is increasingly considered an effective means to reduce the vastness of chemical space into something more manageable for the identification of tool compounds for chemogenomic research and for use as starting points for hit-to-lead optimization. In this section, we highlight the achievements of AI-assisted de novo drug design and point toward potential future developments. The reader is referred to some comprehensive resources regarding de novo molecular design for further information.<sup>454–456</sup>

Early de novo drug design approaches<sup>455</sup> almost exclusively used structure-based methods to grow ligands within the constraints (steric and electronic) of a binding pocket for the target of interest, whether adapted directly from protein structures or inferred from properties of known ligands.<sup>457,458</sup> A limitation of these early methods was that the generated



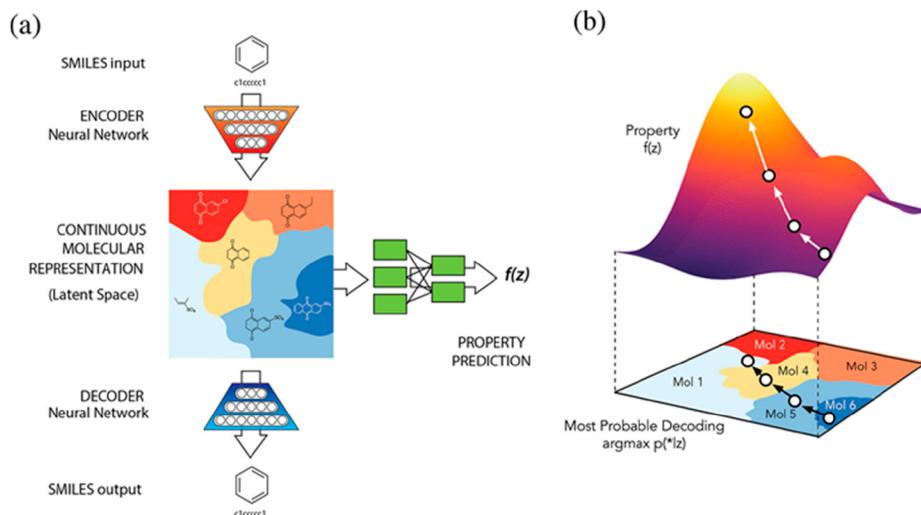
**Figure 22.** Outline of the Bayesian molecular design method, showing the training process for learning important features for a given output variable, the inversion by which regions of the property space are identified as correlating well with the desired output, and the means by which those regions are sampled to suggest novel molecules. Reproduced from ref 464. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

structures were prone to synthetic infeasibility and poor drug-likeness (such as poor drug metabolism and pharmacokinetic properties). More recently, the ligand-based de novo design method has demonstrated its applicability in medicinal chemistry. Generated compound libraries may be additionally analyzed with the aid of a scoring function which takes into account several properties such as biological activity, synthetic accessibility, metabolism, and pharmacokinetic properties.<sup>459,460</sup> One way to build such a virtual library is to use a curated subset of chemical reactions, along with a group of available chemical building blocks, leading to a pool of synthetically accessible molecules.<sup>461</sup> This approach was adopted by Hartenfeller et al.<sup>460</sup> in the development of DOGS, a software which allows the ‘in silico assembly of molecules’ based on a template structure and the aforementioned building block and reaction libraries. Syntheses of these products resulted in novel active compounds against  $\gamma$ -secretase, histamine-4 receptor, and polo-like kinase 1 and in the successful imitation of the pharmacological profile of a natural product, (−)-Englerin A, among other applications.<sup>462</sup> A related approach is to apply knowledge-based expert rules from medicinal chemists to design analogues for a query structure. For example, Besnard et al.<sup>463</sup> used a knowledge-based approach to automatically generate novel dopamine receptor type 2 (DRD2) modulators with specific polypharmacological profiles and suitable ADMET properties for blood–brain barrier penetration. Even though the use of either knowledge-based or reaction rules can reliably and effectively generate novel molecular structures, these approaches are limited by the inherent rigidity imposed by the predetermined rules and reactions. Whether this fact poses a practical problem

may be a matter of discussion. For the purpose of scaffold hopping, however, the cardinalities of the virtual compound libraries that can be generated by a rule-based system easily exceed  $10^{30}$  drug-like molecules.

A third approach, called “inverse QSAR”, deals with the de novo design task from a different angle. Instead of first generating a virtual chemical library and then scoring and ranking it based on similarity to a template compound, inverse QSAR attempts to find an explicit inverse mapping  $y \rightarrow X$  from properties  $y$  to molecular descriptor space  $X$  and then maps back from the favorable region in descriptor space  $X$  to the corresponding molecules.<sup>464–470</sup> The major obstacle of inverse QSAR approaches lies in the selection of a molecular representation which is informative and suitable not only for sufficiently handling the forward QSAR task for a given biological property but also for the subsequent reconstruction stage to be meaningful.

Many de novo drug design methods utilize sets of molecular building blocks or fragments of synthesized compounds for molecule assembly to reduce the risk of generating unfavorable chemical structures.<sup>465,471–473</sup> To avoid overlooking attractive candidate molecules and to increase the diversity and novelty of generated structures, a large fragment library should be used. This comes at the price of a substantial increase in the cost of the fragment swapping and similarity search processes. Ikebata et al.<sup>464</sup> investigated the use of a fragment-free strategy for the generation of novel molecules with desired properties by integrating forward and backward QSAR predictions with the aid of machine learning techniques (Figure 22). They first set up a group of machine learning QSAR models for the prediction of various properties of a given molecule. These



**Figure 23.** (A) An illustration of the autoencoder model for molecular design, including the joint property prediction model. (B) Gradient-based optimization using the GP model in the continuous latent space representation. Reproduced from ref 268. Copyright 2018 American Chemical Society.

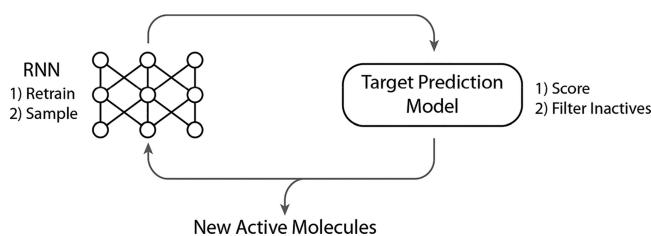
forward models were then inverted through Bayes' law, and the resulting posterior distributions were used to identify high-probability regions for molecules with desired properties. A chemical language model based on SMILES was created to circumvent the problem of chemically unfavorable structures.

As noted by Gómez-Bombarelli et al.,<sup>268</sup> many methods in de novo chemical design depend on explicit rules based on chemical knowledge for replacing or adding molecular fragments to yield new structures, which may bias the search space and ignore certain other structures. Efforts to resolve this issue have primarily focused on means by which to learn such transformations implicitly via generative models.<sup>47</sup> Deep generative networks have shown promise in de novo drug design, without any explicit prior chemical knowledge. Examples of this approach include AAEs,<sup>474,475</sup> VAEs,<sup>268</sup> and RNNs.<sup>476–478</sup> Among the various neural networks, there is a growing interest in RNN-based generative models for the de novo design of molecules,<sup>268,478,479</sup> given their ability to cope with sequential data with long-range dependencies, such as the SMILES chemical representation format and to learn complex grammar. Harel and Radinsky<sup>480</sup> proposed a prototype-driven diversity network, a generative chemistry architecture which combines encoder, VAE, CNN, and RNN components to generate diverse molecules with similar properties to those of a molecular template. They found that the proportion of valid SMILES, i.e., those parsable into sensible molecular structures, from generated suggestions was significantly improved by prototype-conditioning the VAE. This suggests that not all areas of the latent space representation conditioned by the encoder are equally easy to translate into real molecules, likely because of the sparsity in the training data and insufficient penalization of such events during training. Interestingly, 0.01% of the molecules generated using 869 FDA-approved compounds as prototypes were, themselves, FDA-approved.

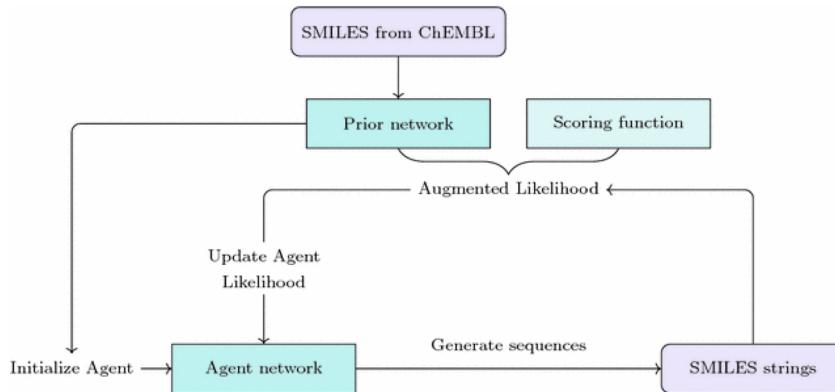
Gómez-Bombarelli et al.<sup>268</sup> proposed a generative model which adopts a somewhat similar strategy. The interconversion between SMILES and continuous latent-space representations was achieved with a VAE and RNN encoder and RNN decoder combination (Figure 23A). To enable molecular design, an additional multilayer perceptron was trained to predict properties of interest based on the latent space coordinates

of molecules. The prediction task was jointly trained on the reconstruction task, so that when given the latent vector of an encoded molecule, new candidate vectors can be generated and decoded into corresponding molecules by moving in the direction most likely to improve the target properties (Figure 23B). The results showed that the model exhibited good predictive power for electronic properties (i.e., orbital energies). Moreover, upon adopting an optimization objective incorporating a mixture of drug-likeness and synthetic accessibility, the model was able to perform iterative, gradient-based optimization to suggest molecules better matching the desired properties (Figure 4 on page 273 in ref 268).

Segler et al.<sup>478</sup> demonstrated that RNNs trained on the SMILES representations of molecules can both learn the grammar required to generate valid SMILES and generate candidate molecules with similar properties to those of template compounds but with differing scaffolds. The de novo drug design cycle of this method adopts transfer learning (Figure 24), in which an RNN model is first trained on a large set of molecules and then further retrained with a small set of active molecules to bias the sampled molecules toward a given template set. Their retrospective results showed that their de novo RNN model could reproduce 28% of 1240 known active compounds against *Plasmodium falciparum*, without having seen the compounds in the initial training, having utilized a



**Figure 24.** Scheme of a de novo design cycle. Molecules are generated based on a chemical language model (RNN) and then scored with the target prediction model. The inactives were filtered out, and the RNN was fine-tuned. Reproduced from ref 478. Copyright 2018 American Chemical Society.



**Figure 25.** An illustration of how the generative model is constructed. Starting from a prior network trained on ChEMBL, an agent is trained using reinforcement learning. Through learning an augmented episodic likelihood which consists of the prior likelihood and a user defined scoring function, the training process aims to fine-tune the prior network toward generating desirable compounds. Reproduced from ref 477. (Creative Commons Attribution 4.0 International License <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

roughly equivalent number for fine-tuning. For *Staphylococcus aureus*, the corresponding figures were 14% of 6051 test molecules, having trained on 1000.

Similar to the work of Segler et al.,<sup>478</sup> Yuan et al.<sup>481</sup> described a new library generation method, Machine-based Identification of Molecules Inside Characterized Space (MIMICS). This method consists of two steps. The first step is to use a character-level RNN (char-RNN) to learn the probability distribution of characters in SMILES strings for given chemical subsets, followed by postprocessing to eliminate structures with invalid valences, aromaticity, or ring-strain issues, resulting in MIMICS output of molecules with similar properties and dissimilar scaffolds to those of the input set. MIMICS-generated compounds were found to act as inhibitors of the unfolded protein response (UPR) and VEGFR2 pathways in cell-based assays, demonstrating the capability of MIMICS to generate useful, novel compounds.

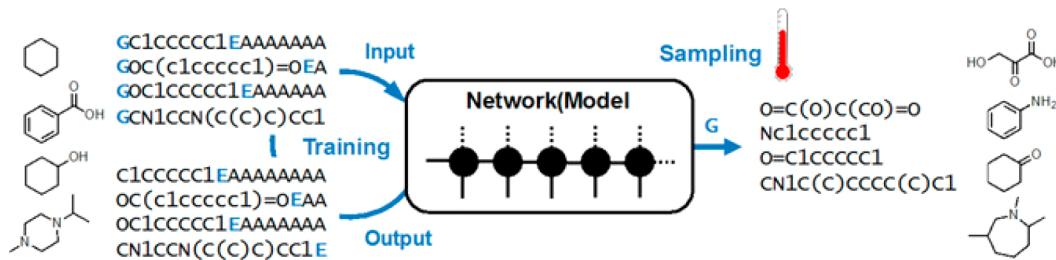
The ability of RNN-based generative approaches to suggest molecules with similar biochemical activities to those of a template or set of templates but with novel scaffolds has been the subject of much interest. For example, Merk et al.<sup>482</sup> developed an LSTM model to generate novel bioactive mimetics of natural products with retinoid X receptor modulating activities. It adopted a transfer learning approach, learning the basic grammar of small molecules from ChEMBL and was then fine-tuned on a small set ( $N = 6$ ) of known natural product retinoid X receptor activators. The generated compounds had a distribution of natural product-likeness<sup>483</sup> scores intermediate between those of ChEMBL and the dictionary of natural products, and 50% of the synthesized molecules showed activity against retinoid X receptor. This approach allows the generation of compounds which incorporate natural-product-like features, while still retaining some of the synthetic feasibility of typical small-molecule compounds and promoting structural diversity. Arús-Pous et al.<sup>484</sup> demonstrated that two-thirds of the GDB-13 chemical space, constructed by means of a rules-based enumeration scheme, can be efficiently reconstructed with an RNN generative model trained on less than one percent of the input space. They found that the model struggled to reproduce more complex structures but performed well overall. The well-defined subset of chemical spaces allowed a rigorous comparison of the model performance to that of an ideal generator (one which only produces valid SMILES, all of

which are part of GDB-13) by using the coupon collector problem to establish the baseline performance.

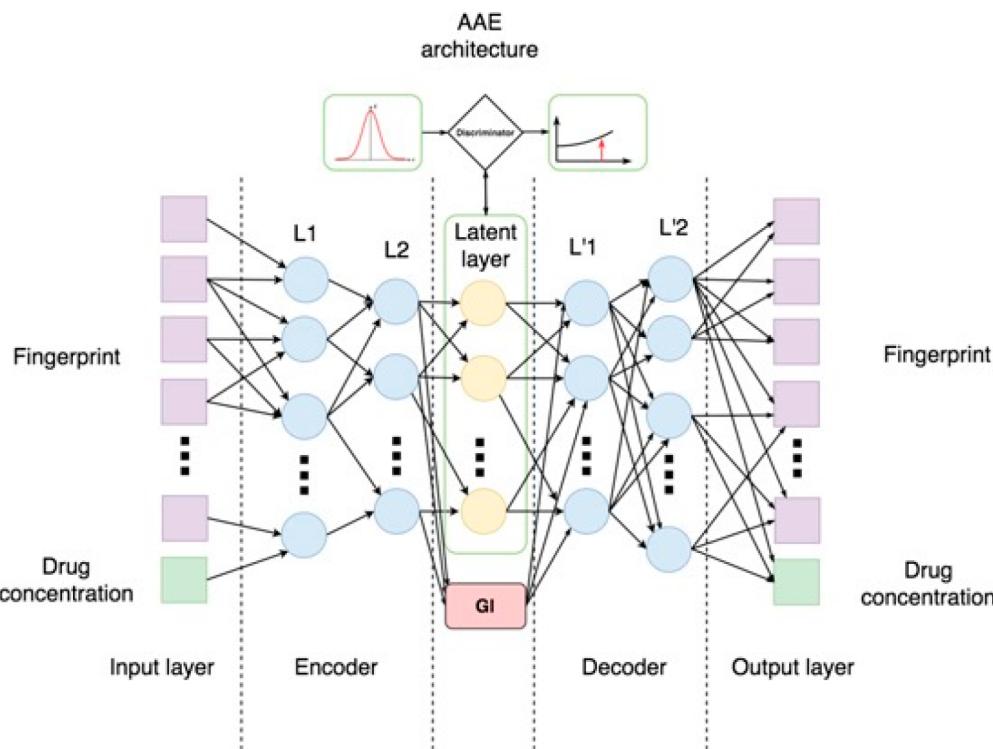
The proportion of valid SMILES in the output of any given generative model is a commonly adopted metric by which to evaluate reconstruction performance. One major contributor to the generation of invalid SMILES is the long-range dependency issue, wherein, for example, the opening and closing brackets representing a ring structure might be separated by many intermediate characters, resulting in an increased likelihood of unclosed rings in the output SMILES. Pogány et al.<sup>485</sup> addressed this issue through use of a bidirectional LSTM architecture coupled with Luong global attention,<sup>260</sup> a strategy which has been observed to improve the performance in terms of long-range dependency sequence generation tasks in other fields. In addition, their approach used reduced-graph representations as intermediate descriptions of molecules and then employed these representations to generate corresponding SMILES which met the pharmacophoric template. Lim et al.<sup>486</sup> proposed a molecular generative model incorporating the conditional variational autoencoder for de novo molecular design. This approach concatenates molecular property information to the latent representation of molecules. The performance of the model was demonstrated by generating drug-like molecules with specific values for five target properties (molecular weight, partition coefficient, number of hydrogen bond donors and acceptors, and topological polar surface area) with a defined margin-of-error and by creating analogues with variable log  $P$  values while constraining the other properties.

There has been some debate concerning the most appropriate means of quantifying the diversity of generated compounds, both internally and with respect to their associated training corpus. Taking inspiration from the analysis of other generative tasks, Preuer et al.<sup>487</sup> suggested the use of the Fréchet distance,<sup>488</sup> a means of comparing distributions, to compare the activation values in the penultimate layer of the ChemNet<sup>489</sup> target prediction for sets of molecules. This approach could be extended to other such molecular representations in a straightforward way and would allow straightforward measurement of the extent to which fine-tuning has biased the sampling toward a given target space, for example.

Similarly, the proportion of unique SMILES generated by any given approach is often viewed as a proxy for its ability to



**Figure 26.** Training and sampling stages of the model by Gupta et al. Reproduced from ref 500. (Creative Commons Attribution-NonCommercial-NoDerivs License: <http://creativecommons.org/licenses/by-nc-nd/4.0/>) Copyright 2018.

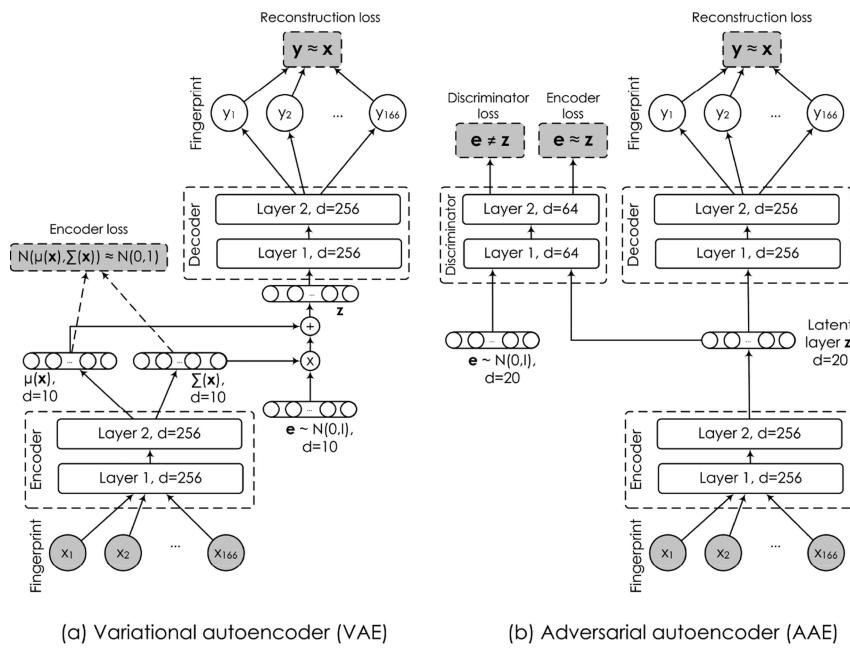


**Figure 27.** Architecture of AAE used for the generation of molecular fingerprints. The encoder consists of two consequent layers, L1 and L2, with 128 and 64 neurons, respectively. In turn, the decoder consists of layers L'1 and L'2, comprising 64 and 128 neurons. The latent layer consists of 5 neurons, one of which is the growth inhibition (GI) percentage, with the other 4 being discriminated via normal distribution. Reproduced from ref 474. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

generate a diverse chemical library. Reinforcement learning<sup>91</sup> has been widely used to ameliorate some of these issues through the fine-tuning of a pretrained generative model, as alluded to above. For example, Jaques et al.<sup>476</sup> recently proposed Sequence Tutor, which improves the generative RNN approach with the addition of reinforcement learning via task-specific rewards, optimizing the approach for  $c \log P$ ,<sup>490</sup> quantitative estimates of drug-likeness (QED),<sup>491</sup> and diversity. The researchers discovered that their reward function tended to result in molecules with relatively simple structures because of their higher probability of both being valid and having the correct property distribution. The difficulty of choosing an appropriate and helpful reward function is well-known.<sup>492–494</sup>

In a study from Olivecrona et al.,<sup>477</sup> a policy-based reinforcement learning method was proposed to extend the previous work<sup>268,478,479</sup> by introducing an “augmented likelihood” approach. In essence, this involves training an agent with estimations of molecule quality under both a prior, pretrained network in chemical space and a scoring function to

guide the model in producing output examples which are still representative of the original chemical domain but reflect the intent of the scoring function employed (Figure 25). They evaluated their generative model in the task of producing molecules without sulfur and found that their model was capable of meeting the scoring objective and creating complex molecules. In addition, they tasked the model to generate molecules similar to a query structure (using the cyclooxygenase inhibitor celecoxib as a template). They found that the agent could generate structurally similar analogues, even when all such compounds were removed from the training set. Furthermore, the model was tested in the task of generating molecules to target dopamine receptor type 2 (DRD2). The researchers found that more than 95% of the generated compounds were predicted to be active based on an SVM model constructed in-group. Some of these structures were experimentally confirmed as active in separate studies but were not included in the training of either the activity prediction or generation models, adding to the activity prediction performed.



**Figure 28.** Schematics of VAE and AAE models. Reproduced from ref 475. Copyright 2017 American Chemical Society.

Reinforcement learning approaches have been applied in other related contexts in drug design, such as in generating ligands against the adenosine A<sub>2A</sub> receptor.<sup>495</sup> Work by Popova<sup>496</sup> et al. considered the extension of the RNN architecture employed in such an approach with persistent memory units, known as stack-RNN.<sup>497</sup> They generated libraries of novel compounds biased toward sets of properties, such as structural complexity, lipophilicity, and inhibitory activity against Janus protein kinase 2.

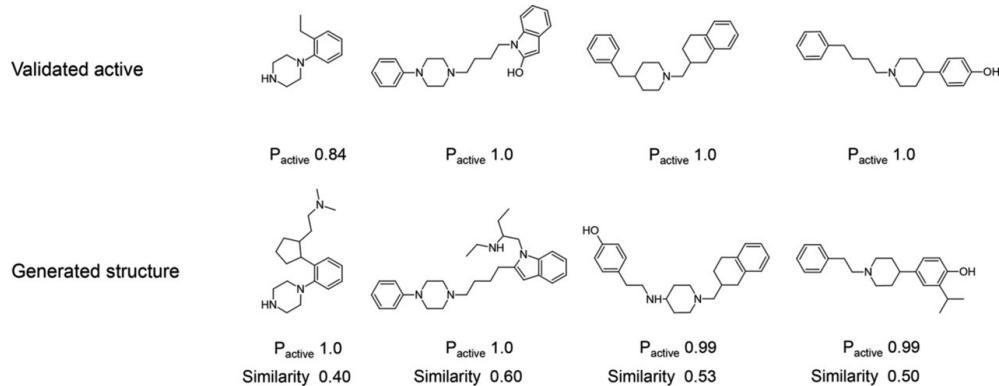
An alternative to reinforcement learning is the transfer learning approach, which aims to improve predictive performance by using insights gained from training on a previous task and transferring them to a new but related task (for more information about transfer learning, see section 2.1.1.6). Awale et al.<sup>498</sup> trained LSTM generative neural networks using molecules taken from commercial catalogs and from FDB-17<sup>499</sup> (a database of fragments up to 17 atoms) and performed transfer learning with ten drug compounds to generate new analogs of these drugs. Their results suggested that transfer learning can learn the rules to assemble small fragments into larger, druglike molecules and that the performance was broadly similar regardless of whether the models were trained on fragments or larger molecules.

Gupta et al.<sup>500</sup> successfully integrated the transfer-learning approach into RNN-based generative models via a new approach to de novo drug design. As shown in Figure 26, an LSTM-based RNN model was first trained on ChEMBL to learn the correct SMILES grammar. The transfer learning technique was used to fine-tune the model to produce SMILES strings which were structurally similar to small libraries of target-focused compounds. Starting from a single receptor-binding fragment, the researchers demonstrated that their generative RNN model could successively grow the remaining molecules. They found that even with a small number of representative molecules used during the fine-tuning process, their approach generated structures with similar chemical characteristics to those of the provided subset (Figure 9 on

page 7 in ref 500), providing a means by which to carry out hit-to-lead optimization with limited data.

The latest class of nonparametric deep generative approaches, GANs, have also demonstrated their ability to generate drug-like molecules with desired properties using SMILES representations as input. Kadurin et al.<sup>474</sup> introduced generative AAEs (sec 2.1.2.6.4) for the generation of molecular fingerprints with definite restrictions in the field of cancer drug discovery. A seven-layer AAE architecture (Figure 27) was designed, including a two-layer encoder for learning how the growth inhibition changed with the molecular fingerprint, a two-layer decoder for reconstruction of the input, and a latent middle layer serving as a discriminator. Particularly, one neuron in the latent layer was in charge of the percentage of growth inhibition of a breast cancer cell line (MCF-7) (called the “GI neuron”), acting as a predictor of compound efficacy against cancer cells. Once the training phase was complete, the sampled latent vectors were decoded to generate fingerprint vectors and a corresponding drug concentration. Those fingerprints with a corresponding predicted drug concentration below 10 μM ( $N = 32$ ) were screened against a library of 72 million molecules from PubChem, and the top 10 most similar hits for each were extracted, resulting in 69 unique compounds. These authors identified a total of 69 novel compounds from this procedure, and while their overall on-target activity is unknown, several of the compounds had previously been identified as anticancer agents. This approach implements a virtual screening process, in which AAE has been shown to be able to identify new molecular fingerprints with predefined anti-cancer cell properties *in vitro*.

In a follow-up study,<sup>475</sup> the above group reported on architecture selection for deep generative models in de novo drug design. They designed an AAE model (Figure 28), called “druGAN” (drug generative adversarial network), for the task of generating new molecules with desired anticancer properties. The paper includes useful information for the practitioner on the design and parametrization of such adversarial architectures, for example, how to calibrate the initial



**Figure 29.** Some generated structures with a high predicted activity value ( $P_{\text{active}}$ ) for dopamine receptor type 2. Their similarity to the nearest validated actives was calculated using the Tanimoto similarity. Reproduced from ref 269. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2018.

performance of the discriminator network such that the generator does not simply reduce coverage of the chemical space. The study showed that when well-parametrized, AAE approaches can be superior to equivalently structured VAE approaches in the achieved reconstruction quality and in the handling of large data sets, at the cost of some loss of coverage of chemical space. They further investigated the efficiency of VAE and AAE generative models as unsupervised feature extractors for the prediction of the aqueous solubility of molecules. Limitations of the proposed architecture include the use of binarized chemical compound descriptor vectors and the requirement of screening the output fingerprints against a chemical library. Using SMILES directly would allow a more intuitive interpretation and permit the end-user to evaluate whether or not the model is making sensible predictions in a more straightforward manner. The developers suggested that their druGAN model allows new molecules to be proposed for a variety of targets, allowing end-users to narrow their search space.

The entangled conditional adversarial autoencoder proposed by Polykovskiy et al.<sup>501</sup> is another novel AAE model which improves upon the original supervised adversarial autoencoder architecture,<sup>242</sup> allowing one to incorporate target variables with a more rigorous “disentanglement” of the target ( $y$ ) and latent ( $z$ ) variables, which is, in essence, a complexity- and redundancy-elimination procedure. The original method simply concatenates normalized target variables to the latent vector representation of a molecule, as with the work of Lim et al.<sup>486</sup> In the study, the researchers propose two approaches, with the first being predictive disentanglement, which uses an ANN model to learn  $p(y|z)$  and remove features which have substantial predictive power for  $y$  from  $z$ , both of which represent redundant variables. The second approach is joint disentanglement, which, in brief, trains the network to distinguish between real pairs of  $z$  and  $y$  variables and between pairs of noise ( $\epsilon$ ) and  $y$  variables. Molecular structures were generated with various target properties, such as activity against a specific protein, solubility, or ease of synthesis. The authors applied the proposed model to select a promising hit compound active toward Janus kinase 3, implicated in rheumatoid arthritis, psoriasis, and vitiligo, from a pool of 300,000 generated compounds by a combination of virtual screening, molecular dynamics, and expert opinion. This molecule was synthesized and tested in vitro, showing moderate activity and subtype selectivity, demonstrating the

potential for integrating generative adversarial models in drug discovery pipelines.

Méndez-Lucio<sup>502</sup> reported a new method based on the stacking of conditional generative adversarial networks. In this method, the researchers generated hit-like molecules from gene expression signature-perturbagen structure pairs released as part of the CMAP project.<sup>503</sup> They found that the generated compounds had higher similarity to known active compounds than can be accomplished by selecting compounds based on comparisons of expression signatures for a pool of withheld data.

Blaschke et al.<sup>269</sup> compared the performance of adversarial autoencoders as structure generators with several VAE instances. The VAE instances were named teacher and no-teacher VAE, with “teacher” referring to teacher-forcing, a technique in which a model gives both the output of the model and the corresponding character from the training corpus as inputs for the next time-step. Molecular structures were encoded into a continuous latent space and then decoded into the original space to determine the loss introduced by this compression process. The reconstruction accuracy of all models was at least 95%, with an observed benefit of the adoption of teacher-forcing in the VAE case during training, with the effect inverted during generation. The overall reconstruction accuracy was higher for all AAE models considered and best of all for the uniform AAE, which forces a uniform distribution onto the latent vectors. If the latent space is well constructed then the distance between compounds in their original space should be preserved. This behavior was used to sample latent space vectors at increasing distances from the vector corresponding to celecoxib, leading to analogous compounds being proposed, and to confirmation that the distance in latent space corresponds with the Soergel distance between the compounds’ extended-connectivity fingerprint 6 (ECFP6) representations.

By searching for new compounds in the latent space using a Bayesian optimization process, new structures predicted to be active against dopamine receptor type 2 by a QSAR model were identified (Figure 29).

Attempts to combine adversarial and reinforcement learning strategies to suggest molecules which are appealing from a medicinal chemistry perspective are ongoing. The ORGANIC (objective-reinforced generative adversarial network for inverse-design chemistry) architecture was proposed by Sanchez-Lengeling and colleagues.<sup>504</sup> ORGANIC combines

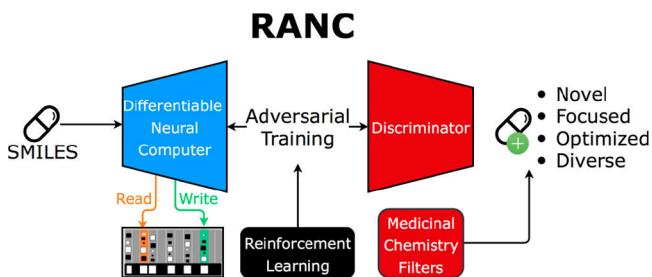
GANs and reinforcement-learning as follows. The generator is given the task of producing molecules which can defeat the discriminator, thus biasing its output toward the initial distribution, and is rewarded by the reinforcement process, fulfilling the desired design outcome. The relative importance of these two characteristics is tunable by the end-user. In slightly more detail, the reinforcement element is included in the generation process by setting an action space equivalent to the space of all sampleable characters and performing a Monte Carlo search with a canonical rollout policy to determine the long-term benefit of possible next-character choices, given the current generated sequence of characters. Surprisingly, the authors found that the model, although well-performing in some predictive aspects, struggled to generate valid SMILES on a consistent basis, with different experiments producing between 0.1 and 99.98% valid SMILES.

The original ORGANIC architecture suffers from several common problems found in adversarial settings such as mode-collapse during training. In an effort to study and improve the convergence properties of GANs, Aspuru-Gudzik's and Zhavoronkov's groups proposed the reinforced adversarial neural computer (RANC) method,<sup>505</sup> extending the ORGANIC paradigm (Figure 30) for the de novo design of novel small-

addition to introducing the popular differential neural computer technique developed by DeepMind,<sup>506</sup> the adversarial threshold neural computer model has a unit called the “adversarial threshold”, which acts as a filter between the agent (generator) and the environment (discriminator + objective reward functions). This approach produced similar increases in the validity and tuning of chemically interesting properties to those achieved by the RANC approach discussed above but showed a smaller improvement in diversity. The efficacy of the adversarial threshold neural computer model at producing hit compounds was confirmed by a preliminary biological evaluation of the molecules generated by the adversarial threshold neural computer model, leading to the discovery of seven kinase inhibitors with novel scaffolds from a pool of 50 tested compounds.

The works described above demonstrate the effectiveness of SMILES-based models in molecule generation. However, these models suffer from some problems. First, the proportion of valid SMILES produced by the models varies substantially, and there is no guarantee that they will represent a reasonable molecular structure. Several routes to alleviating this problem have been proposed, such as using a grammatical definition of the SMILES syntax<sup>509,510</sup> or transforming a SMILES string into another representation which conveys identical information but is more suited to machine-learning.<sup>511</sup> However, these approaches do not necessarily address the underlying issue of long-range dependencies and multiple valid representations in the reconstruction of valid and relevant SMILES.

Second, a single-character perturbation in a text-based representation (e.g., a SMILES representation) of a molecule can result in a significant change of the underlying molecular structure or even render it invalid.<sup>512</sup> In addition, partially generated text representations, in many cases, are meaningless. As an alternative strategy to replace SMILES representation, the development of generative models working directly on the molecular graph has recently become a popular research topic. Molecules directly represented as molecular graphs are somewhat more robust than are text sequence representations. Furthermore, partially generated molecular graphs can be interpreted as substructures, such that researchers can perform chemical checks, e.g., of the atom valency, on a partially generated molecule. Most existing graph generators need to train either a GAN or a VAE architecture, wherein specific designs are often needed for different applications. Recent graph-generative models with a GAN structure include MolGAN<sup>513</sup> and Mol-CycleGAN.<sup>514</sup> MolGAN predicts discrete graph structures in a single step (i.e., nonsequentially). Internally, the generator produces a dense adjacency tensor, and an annotation matrix, which it then categorically samples to create a more meaningful molecular representation. Finally, it utilizes a permutation-invariant discriminator to solve the node variant problem in the adjacency matrix, generating an output molecule. A reward network is introduced in parallel into MolGAN to bias the optimization toward desired chemical properties. Although this model produces a high proportion of valid molecules and makes considerable progress toward solving some of the issues involved in working with molecular graph representations, it faces limitations in choosing to learn an implicit distribution rather than an explicit one. Furthermore, experiments have proved that it also suffers from the well-known “mode collapse” problem in generative adversarial networks, in which a very small section of the sample space is continually reproduced.



**Figure 30.** Architecture of the reinforced adversarial neural computer (RANC) model. The generator is a differentiable neural computer which uses differentiable external memory and attention mechanisms. After being pretrained with the maximum likelihood estimation, the generator is trained with reinforcement learning and adversarial training to read and write sequence structures that are relevant for the reward. Reproduced from ref 505. Copyright 2018 American Chemical Society.

molecule organic structures. RANC uses a differentiable neural computer<sup>506</sup> as a generator in place of the original LSTM, which contains a memory bank to allow the capture of long-term dependencies, as well as providing generative capacity, to balance the generator and discriminator during adversarial training. In essence, this is a similar approach to that proposed by Popova et al. but with different architectures. It increases the average proportion of valid SMILES generated, from 7% to approximately 70%. RANC improves upon the performance of ORGANIC<sup>504</sup> under several metrics of importance to drug discovery, doubling the proportion of unique structures and doubling those which pass medicinal chemistry filters and the Muegge<sup>507</sup> criteria for the selection of druglike compounds, as well as achieving higher quantitative estimates of drug-likeness scores overall. Furthermore, RANC is capable of generating structures that match the distributions of key chemical features and descriptors in the training data set, such as the length of SMILES, molecular weight, log P, TPSA, etc.

Another extension of the ORGANIC paradigm is the adversarial threshold neural computer architecture for the de novo design of novel small-molecule organic structures.<sup>508</sup> In

Mol-CycleGAN is a CycleGAN-based generative model<sup>515</sup> that operates in the latent space trained by JT-VAE,<sup>516</sup> a VAE trained on junction tree molecular representations rather than SMILES, with the purpose of generating molecules with the desired properties while retaining their chemical scaffolds. They demonstrated that 100% of samples resulted in a valid molecule, and that they could optimize the results for  $c \log P$ . Recent graph generative models with a VAE structure include GraphVAE,<sup>517</sup> JT-VAE,<sup>516</sup> and several probabilistic iterative graph generation models.<sup>518</sup> GraphVAE is designed to directly generate the adjacency matrix of molecular graphs, as well as the vertex and edge features. However, the generation process used in the GraphVAE model has proven computationally expensive and can only be applied to the smallest molecular graph structures (approximately 10 atoms).

Instead of generating molecule sequences atom-by-atom, which is likely to result in invalid intermediate molecules, JT-VAE defines another, more chemically rooted generation process in which molecules are represented both as graph structures and as a junction tree of small clusters of atoms, roughly corresponding to molecular substructures. In the generation phase, it produces a junction tree node-by-node in a depth-first search, evaluating at each point whether any given node should have children and, if so, adding them and predicting their substructure labels. After this process is complete, the substructures matching each substructure label are scored and placed, resulting in a complete molecule.

Liu et al.<sup>519</sup> also recently proposed a constrained graph VAE model. The decoding process begins by initializing a set of node latent variables and then performing steps of edge generation, edge labeling, and node updating until a special termination node has been selected in the edge generation step. This process is repeated for a new node if there are no valid candidates in the connected graph. As part of this process, valency rules are used to help ensure that syntactically valid molecules are constructed, along with other masks designed to incorporate chemical logic. As with other graph-based models, 100% of sampling efforts result in a valid molecule, with 95–100% of these being novel, and approximately 100% being unique. As an extension, the authors demonstrated that they could train a quantitative estimate of drug-likeness predictor on their latent space and use this predictor to guide a sampling trajectory toward higher quantitative estimates of drug-likeness-scoring generated molecules.

To the best of our knowledge, no rigorous comparison has been performed to date for the various molecular graph-based approaches that would allow one to estimate the overall utility of the approaches and their ability to generate novel chemical matter. Evidently, only by synthesizing and testing de novo designed compounds will we be able to assess the various techniques.<sup>261</sup> Other approaches to the generation of molecular graphs have been adopted,<sup>520,521</sup> but the examples introduced herein describe the major trends in current research and efforts expended to date. Overall, these methods are not without issue. Training models to generate sensible molecules which obey chemical logic, are novel, and satisfy certain user-expectations remains a challenging task.<sup>521</sup> Nevertheless, the pace of progress and the incorporation of novel ideas from related fields suggest that the combinations of architectures being investigated could help with tasks such as the generation of analogues with a specific scaffold, promoting drug-likeness and synthetic accessibility, and rationally designing polypharmacological agents.

AI-assisted de novo drug design<sup>496,522,523</sup> can provide a broader exploration of chemical space than was possible with blunter tools and has shown its potential value for novel ligand scaffold identification. Despite the newly developed AI algorithms for fully automated molecular de novo design, the human task of picking the most promising candidates has proven helpful in beginning the optimization stage. It is unlikely that such approaches will generate only synthetically feasible, active compounds with ideal ADMET properties in the near future, although complex objective functions incorporating information such as this are under development. For now, one can expect that such methods should propose novel compounds with a reasonable likelihood of synthetic success and that are inspiring candidates for new lead series to fuel hit-to-lead optimization.<sup>524</sup>

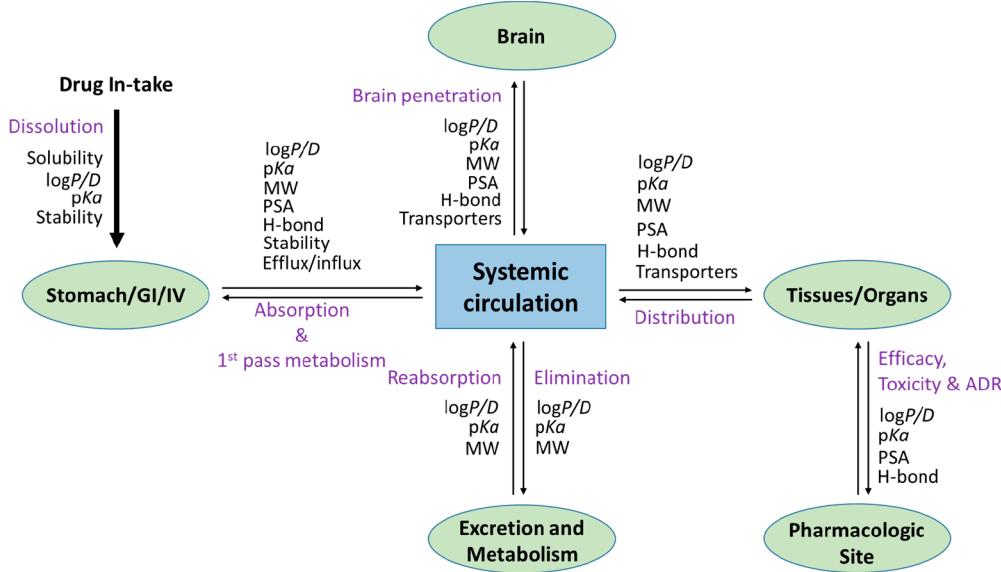
To achieve further acceptance of compounds generated from AI techniques, it will be important to understand the current limitations of this design approach. One of the main limitations is the lack of informative and suitable ways to translate the domain knowledge learned by the AI model into molecular structures. Even though there are various choices for molecular representation, such as SMILES strings, molecular fingerprints, descriptors, or novel representations based on chemical graphs, the most popular method to-date in AI-assisted de novo drug design is the SMILES representation of two-dimensional molecular graphs because of the ease of conversion to a molecular structure. There is still room for improvement of these outputs in terms of validity and chemical novelty.

Nearly all generative models in the field of AI-based de novo drug design barely consider the structural (steric and electronic) information on the target protein, although examples of deep learning networks for predicting protein–ligand binding affinity that incorporate protein structural information exist.<sup>326</sup> The most frequently used trick is to leverage the information about the activity of small molecules with an additional adversarial approach and/or reinforcement learning approach to guide the generative process toward certain desired criteria.

The lack of standardized approaches for the analysis and scoring of generated molecules renders it difficult to create an honest appraisal of the merits and pitfalls of each approach. Proxy measures, such as the proportions of valid and unique SMILES, give an overview of the generative capacity but little insight into the representative aspect. For more details of the achievements and remaining challenges in this field, see section 3.2.<sup>525</sup> Next-generation small-molecule drugs will be designed to interact with multiple targets.<sup>526,527</sup> This multidimensional view of chemical and biological ensembles is a perfect ground for AI-based de novo drug design, as the incorporation of multiple objectives can improve the power of such systems in a similar vein to that of the multitask learning approaches discussed earlier.

## 6. PHYSICOCHEMICAL AND ADMET PROPERTIES PREDICTIONS

Improperly balanced ADMET properties are a common source of late-stage failure of drug candidates and have led to the withdrawal of approved drugs.<sup>528,529</sup> The importance of determining the ADMET properties of a compound as early as possible in the drug discovery process has been widely recognized.<sup>530,531</sup> Therefore, the development of *in silico* models has attracted great interest from pharmaceutical companies and academic groups, as such models can provide



**Figure 31.** Schematic representation of the relationship and interplay of physicochemical properties with in vivo drug pharmacokinetics/pharmacodynamics. Adapted with permission from ref 545. Copyright 2009 Verlag Helvetica Chimica Acta AG, Zürich.

valuable information to guide the synthesis or construction of a screening library that minimizes such issues.

Over the past decades, AI techniques have been widely used to directly correlate the ADMET properties of chemicals with molecular descriptors (or features) and to construct predictive models from available data sets. With an increasing understanding of the relationship between the bulk physicochemical properties [such as the octanol–water partition coefficient ( $\log P$ ), aqueous solubility ( $\log S$ ), and intrinsic permeability] and the biological behaviors of a compound, particularly its pharmacokinetic properties, indirect methods for rapidly predicting ADMET properties have been developed. These approaches are designed to guide drug discovery efforts to potentially more productive areas of chemical space, without placing harsh constraints on specific chemical groups at this stage. This section focuses on the strategies adopted, progress made, and remaining difficulties in the prediction of physicochemical and ADMET properties between 2015 and the present day. Earlier advances in this area, as well as other in silico models developed to predict physicochemical properties and ADMET properties, are extensively described in the literature.<sup>531–536</sup>

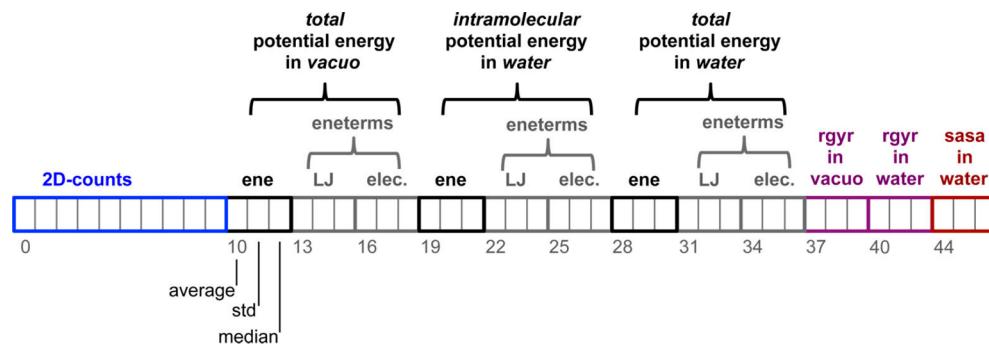
### 6.1. Prediction of Physicochemical Properties

The simple physicochemical properties of a small-molecule drug candidate have a well-known impact<sup>537,538</sup> on its eventual success in entering the market through influences on the ADMET properties (Figure 31), potency, and selectivity of the drug compound.<sup>539–543</sup> For example, familiar physicochemical properties of compounds have been used to classify their likely target family (e.g., nuclear receptor ligands have mean  $\log P$  and molecular weight values of 6.9 and 482 Da, respectively, whereas the mean corresponding values for G-protein-coupled receptor ligands are 4.8 and 573 Da<sup>544</sup>). This method is likely confounded to some extent by the development of small molecule classes and me-too approaches, but it is still an illuminating insight. Small molecule drug candidates must display sufficient solubility and permeability to reach their site of action and engage their targets. Therefore, a better understanding and accurate prediction of the physicochemical

properties should be beneficial for the design of compounds with desired pharmacokinetic and pharmacodynamic profiles.

**6.1.1. Lipophilicity Prediction.** As one of the most important physicochemical properties for drug discovery efforts, lipophilicity plays a crucial role in modulating many key pharmacokinetic processes.<sup>533,546–548</sup> Specifically, it affects the membrane permeability of drug molecules and therefore impacts drug transport, distribution, and clearance behaviors. It also strongly influences the binding of drug molecules to macromolecules, affecting metabolic and toxic processes, as well as almost any other drug-induced biological process in the body. The gold standard for the quantitative characterization of lipophilicity is the logarithm of the octanol–water partition coefficient ( $\log P$ ) or pH-dependent distribution coefficient ( $\log D$ ); alternatives include liposome/water partitioning and immobilized artificial membranes (IAM) chromatography. Conventional computational approaches for predicting lipophilicity include group-contribution methods (GC),<sup>549</sup> equations of state,<sup>550</sup> quantum chemistry-driven methods such as conductor-like screening model (COSMO)/conductor-like screening model for real solvents (COSMO-RS),<sup>551</sup> molecular simulation,<sup>552</sup> and linear/nonlinear QSPR.<sup>553</sup> For each of these approaches, researchers have increasingly commonly adopted an AI-driven approach, correlating  $\log P/\log D$  with molecular descriptor sets. A large number of published studies<sup>554–557</sup> have utilized this protocol, and the commonly employed molecular descriptors include the atomic charge, hydrogen bonding effects, molecular volumes, and surface areas. For example, Zang et al.<sup>558</sup> adopted a similar protocol in 2017 for an open-source workflow to predict six physicochemical properties, including  $\log P$ ,  $\log S$ , melting point (MP), boiling point (BP), vapor pressure ( $\log VP$ ), and bioconcentration factor (BCF).

Neural networks have been utilized to predict the octanol–water partition coefficient, such as the ADMET Predictor and the ALOGPS program, to name a few. In particular, the ALOGPS program, based on an associative neural network<sup>559</sup> that combines elements of the feed-forward network and the kNN approach, has been proven to reliably predict  $\log P$  values



**Figure 32.** Schematic representation of MDfp+ which includes two-dimensional counts and properties extracted from MD trajectories: the radius of gyration (rgyr), solvent-accessible surface area (sasa), and the intramolecular and total potential energy of the solute (ene) together with its Lennard-Jones and electrostatic terms (eneterms). Each property is represented by its average, standard deviation, and median. Reproduced from ref 570. Copyright 2017 American Chemical Society.

for low molecular weight compounds<sup>557,560,561</sup> and has been applied by several groups.<sup>562–566</sup>

Two types of descriptors have typically been used to build predictive models, structural fragments, and numeric indices, from which useful subsets are usually selected using a genetic approach.<sup>558,567</sup> The limited molecular-weight range of existing models and the likely correlation with the properties captured by the descriptors chosen has led to calls for improvement of log *P* predictions by including more physically meaningful representations (e.g., quantum chemical descriptors)<sup>568</sup> or by integrating average properties obtained from molecular simulations (MC or MD simulations) into the molecular descriptors.<sup>569</sup> For example, a novel MD feature representation, termed MDfp+ (Figure 32) was introduced by Riniker in 2017 to predict log *P*.<sup>570</sup> The author found that their information-rich fingerprint achieved comparable performance with more rigorous MD-based calculations. This approach might lead to the development of a more rigorous and extensible class of log *P* models, although the computational load is a significant hindrance.

AI-based protocols have been introduced based on octanol–water as the reference system to reliably predict the log *P* (or log *D*) with a large variety of diverse data sets and thus to support the selection of the most promising compounds for the initial investigation. The main difficulty in predicting log *P* (or log *D*) is related to aqueous solubility. These AI-based protocols vary in accuracy and efficiency but are all trained on experimental data, which may limit the domain of applicability compared to other physical-based approaches such as conductor-like screening model for real solvents or molecular simulation.<sup>571</sup> Importantly, relatively few groups have rigorously defined the domain of applicability for their models. A recent study by Fuchs et al.<sup>572</sup> highlighted the importance of this concept, with several commercially available softwares being compared on their ability to generalize to other data sets.

**6.1.2. Aqueous Solubility Prediction.** Sufficient aqueous solubility (expressed as log units of the molar solubility ( $\text{mol L}^{-1}$ ), log *S*) is crucial for orally administered drugs, since they must first be dissolved in the gastrointestinal fluids before absorption from the gastrointestinal tract (Figure 31).<sup>573</sup> Numerous in silico approaches have been proposed to predict aqueous solubility over the years and are grouped into three broad categories based on their founding principle: quantum chemical approaches with varying theory levels,<sup>574</sup> molecular mechanics simulations,<sup>575</sup> and descriptor-based approaches founded in statistics or machine learning algorithms. In this

section, we focus on some of the most recent developments in predicting the log *S* value of small molecules with deep neural networks. For additional details on other computational approaches designed to predict aqueous solubility, we refer the reader to comprehensive review articles.<sup>576–583</sup>

The prediction of aqueous solubility is often based on four major descriptor groups, incorporating information about (i) the physicochemical properties, (ii) atomic or group contributions, (iii) quantum chemical properties, and (iv) structure. More specifically, physicochemical property-based descriptors relate to properties that can be measured experimentally or obtained from computational approaches, e.g., log *P*, and the melting point ( $T_m$ ).<sup>584–587</sup> The general solubility equation (GSE) method, with its extended and modified variants, is another example of a physicochemical-type method that estimates log *S* from log *P* and  $T_m$ .<sup>588</sup> Although this method produces good results, it adds additional complexity by replacing the problem of predicting aqueous solubility with those of measuring or predicting both log *P* and  $T_m$ , which are nontrivial, as discussed above.

Approaches using atomic or group contribution descriptors capture the presence of atoms or groups in a molecule and estimate the contribution of each to the log *S* through correlation with available experimental data.<sup>589,590</sup>

Structure-based methods predict solubility based on chemical structural information, such as molecular topology, connectivity, or fragment information.<sup>591</sup> In addition to the molecular size, a key feature that is traced back to the work of Führner,<sup>592</sup> other molecular features related to log *S* have also been discovered over the years, including the number of hydrogen bonds<sup>593</sup> and various molecular connectivity indices.<sup>594,595</sup>

Aspects of molecular structure have been represented in several ways,<sup>267,596,597</sup> with a typical choice being a binary fingerprint that offers a good trade-off between simplicity and predictive power.<sup>598–600</sup> Examples of fingerprints used to predict aqueous solubility include the extended connectivity fingerprint 2<sup>597</sup> which encodes each atom and its molecular environment within a diameter of two chemical bonds,<sup>598</sup> and neural fingerprints, which are generated directly from molecular graphs of an arbitrary size and shape.<sup>267</sup>

Several deep architectures and deep learning approaches have also been used to predict the aqueous solubility of druglike molecules, such as undirected graph recursive neural networks (UG-RNNs)<sup>601</sup> and graph-based convolutional neural networks.<sup>602</sup> These approaches focus on learning

specific information about local environments with the aim of constructing task-specific internal representations. The former approach, which was described by Lusci et al.,<sup>601</sup> learns descriptor representations by enumerating directed acyclic graph representations of molecules. The latter approach, which was proposed by Coley et al.,<sup>602</sup> convolves over molecular graphs to construct a type of task-specific Morgan fingerprint and showed a significant improvement over other methods for a solubility task, as well as octanol solubility and  $T_m$  tasks. In each case, relatively few molecules (approximately 1100) were available for model building and testing.

Although many informative descriptors have been designed and incorporated into the construction of AI predictors, none seem capable of predicting solubility with perfect accuracy.<sup>603</sup> In an assessment of the quality of different predictors based on four different approaches (multiple linear regression, ANNs, category formation, and four commercially available models), Hewitt et al.<sup>603</sup> found that none perfectly predicted the solubility, although multiple linear regression performed better than other, more complex approaches due to its lower probability of overfitting. In addition to an insufficient appreciation of the inherent complexity of the property, another frequently cited reason for the limited predictive ability of current models is the noisiness of solubility data collected from multiple sources in the literature, with estimated root-mean-square errors (RMSEs) in experimental data ranging from 0.6–0.7 log  $S$  units.<sup>580,604</sup>

However, a recent investigation by Palmer and Mitchell has indicated that the quality of the experimental data was not an overly limiting factor in predicting the aqueous solubility of druglike molecules.<sup>605</sup> The authors utilized a CheqSol solubility data set<sup>604,606</sup> that contained the intrinsic aqueous solubilities of 132 drug molecules with high reproducibility and approximately 0.05 log  $S$  units of standard error. As a comparator, they also employed a much noisier data set that was extracted from several different sources from the published literature with 0.6–0.7 log  $S$  units of uncertainty. In contrast to their expectations, the models derived from the CheqSol experimental data were not more accurate than those derived from published data. Therefore, they considered the resulting RMSE an indication of a deficiency in QSPR methods (algorithms and/or descriptor sets) rather than a product of uncertainty in the experimental measurements, as commonly suggested. The contribution of the data set size was accounted for in this study, but the authors state that the estimation of the importance of diversity in the subset was not considered.

Overall, although aqueous solubility is an important physicochemical property of small molecule drugs and drug candidates, it remains frustratingly difficult to predict accurately.<sup>603</sup> Each prediction approach has its own benefits and limitations, and the best choice remains largely dependent on individual requirements. For high-throughput molecular discovery, efficient AI protocols represent a suitable and sufficiently reliable approach for a broad-brush refinement of large sets. However, when a specific physical and mechanistic understanding of the underlying principles of a few specific drug molecule(s) is desired, researchers typically prefer to use more physically meaningful methods based on quantum mechanics and/or molecular mechanics.

**6.1.3. Intrinsic Permeability Prediction.** An evaluation of in vitro membrane permeability data is crucial for drug discovery and development efforts based on the importance of the permeation of molecules across biological membranes via

passive transmembrane diffusion and/or active transport mechanisms in most pharmaceutically relevant biological processes.<sup>607–609</sup> Significant efforts have been expended to develop computational models that accurately predict the membrane permeation behavior of small molecules.<sup>610–613</sup> For this purpose, many AI-assisted predictors have been trained with diverse chemical classes and their associated experimental permeability data. These data have been collected in various in vitro permeability assays, including Madin-Darby canine kidney (MDCK) cells, the parallel artificial membrane permeability assay (PAMPA), the human colon adenocarcinoma (Caco-2) cell line, and other cell-based assays.

The MDCK cell line is increasingly being used in drug discovery processes to assess the membrane permeability of a compound. However, the presence of efflux transporters, such as P-glycoprotein 1 (MDR1), poses an issue for the estimation of intrinsic permeability.<sup>614</sup> Therefore, special caution is required when collecting the initial data set. One possible means of addressing this issue is to exclude all compounds that are known substrates for efflux transporters from the training set, as described by Broccatelli et al.<sup>615</sup> The authors suggested that the discovery process might benefit substantially from an early, judicious application of in silico models for the prediction of permeability, as well as from a careful analysis of the observed disparity between in silico and in vivo values, describing the partial least-squares and SVM models for MDCK cells that have been used at Genentech to filter out compounds prior to synthesis.

The Caco-2 cell model is a well-established drug discovery screening tool in research institutions and pharmaceutical industries<sup>616–618</sup> and has attained a central position in drug permeability screening in industry. The Caco-2 cell assay has been extended to achieve a Caco-2 cell-based intrinsic permeability assay by studying compounds at multiple concentrations to saturate active transport<sup>619</sup> or by adding chemical inhibitors of important efflux transport proteins. For example, Fredlund et al.<sup>620</sup> recently reported the validation of a high-throughput assay designed to measure the apparent intrinsic permeability ( $P_{app}$ ) in the presence of effective chemical inhibitors against the three major efflux transporters P-gp, MRP2, and BCRP in the Caco-2 cell line. Machine learning models have been developed using their in-house AutoQSAR<sup>621</sup> system to correlate the intrinsic permeability of Caco-2 cells with molecular structure. Because the performance of prediction models may vary for specific projects, the authors also proposed a “permeability decision tree” to help guide project design by combining the predicted permeability with lipophilicity and molecular weight.

Intrinsic permeability has also been readily assessed using PAMPA.<sup>622</sup> PAMPA is a noncell-based method with several advantages, including a low cost, good time effectiveness, higher tolerance to a wider pH range and higher DMSO content than cell-based approaches, and an amenability to high-throughput screens.<sup>623,624</sup> When building models for predicting PAMPA permeability, a relatively large permeability data set is available,<sup>625</sup> which includes more than 5435 structurally diverse compound entries generated under the same assay conditions.

Care must be taken when collecting the data set(s). Only permeability data measured under consistent conditions can be collected, unless condition information is integrated in the model-training process. At the same time, the influence of different transporters expressed in cell lines, as well as the

demand for diverse compound libraries should be considered. Recent publications have reported the power of machine learning methods in predicting intrinsic permeability. These investigations clearly indicate that models based on machine learning methods can and should be used in the early stages of drug development to filter out poorly permeable compounds with a high degree of confidence and to focus synthetic efforts on regions of chemical space with the potential for optimal permeability.

**6.1.4. Predictions of Other Physicochemical Properties.** In addition to the three important physicochemical factors already mentioned above, other properties, including, but not limited to, the ionization constant and  $T_m$ , which relate to ADMET events in the body, are therefore important for efficacy, safety, and metabolism. Many AI-based tools have been developed and deployed to model physicochemical data.<sup>626–629</sup> Rupp et al.<sup>630</sup> used graph kernels and kernel ridge regression to describe molecules based on their molecular graphs and constructed a model to estimate acid dissociation constants, achieving an accuracy similar to a semiempirical model developed by Tehan et al.<sup>631,632</sup> Zhou et al.<sup>633</sup> developed an ANN-based group contribution model to estimate the  $pK_a$  with an average absolute error of 0.17. Recently, Li et al.<sup>634</sup> introduced an improved particle swarm optimization algorithm and then used it to select features for an RBF ANN trained to predict the  $pK_a$  values of 74 structurally diverse neutral and basic molecules. Validation with a data set consisting of 20 molecules yielded a good prediction performance with an RMSE of 0.04.

Overall, physicochemical properties have a crucial influence on the ADMET characteristics of a compound, with a strong interdependence between factors. A typical example is that an increase in lipophilicity enhances blood-brain barrier permeability but may also negatively impact blood clearance and increase plasma protein binding, resulting in an overall decrease in the free drug concentration in the brain.<sup>635</sup> Therefore, balancing multiple physicochemical parameters and improving one property without negatively affecting another have become key challenges for the medicinal chemist in drug discovery and development. This mutual interdependence of properties and the inherent complexity of the factors influencing the disposition of the compounds *in vivo* create challenges in correlating physicochemical properties with ADMET profiles and developing reliable AI-based predictors. Multiobjective approaches are becoming increasingly feasible, due to the quantity and variety of publicly available data, and the development of architectures for which such objective functions are straightforward, such as the combination of GAN and reinforcement learning methods, is discussed in section 5. This approach promises a more holistic understanding of drug-body interactions and a means to rationalize optimization efforts in a reasonable time frame.

## 6.2. ADMET Predictions

Early ADMET prediction help researchers select good drug candidates<sup>636,637</sup> and promote drug-likeness throughout the development pathway.<sup>636,637</sup> Recent years have witnessed many advances in the field of AI-assisted ADMET predictions, and a new generation of predictors based on deep learning and big data will benefit the process of drug discovery and development from lab-to-clinic.

**6.2.1. Absorption.** Drug absorption is the crucial first step in which drug molecules are absorbed into the bloodstream

from the absorption site through one or more biological membrane barriers. It is a complex process that is substantially influenced by several physicochemical properties of drugs, along with the route of administration, the formulation chosen, and physiological factors. Predictive models for drug absorption are divided into two major domains, physicochemical and physiological. Currently, several simple filter approaches based on evaluating physicochemical properties involved in the absorption process, including, but not limited to,  $\log P$ ,  $\log S$ , and intrinsic permeability, have been proposed for a quick prediction of the bioavailability of small molecules (for a more detailed description of these physicochemical properties, see section 6.1).<sup>638,639</sup> These approaches help medicinal chemists make preliminary and rapid judgments in hit-to-lead optimization processes, e.g., a drug candidate molecule is expected to have poor absorption if it has low intrinsic aqueous solubility. However, the prediction of drug absorption based solely on physicochemical features struggles to achieve reasonable accuracy, due to the multiple drug absorption mechanisms across a biological membrane, and every missed molecule is a potential blockbuster. Therefore, a physiological model assists in the development of a reliable, high-throughput method.

Human intestinal absorption (HIA) is one the most influential ADMET properties and directly affects oral bioavailability. A number of computational classification and regression models have been introduced to predict HIA based on data from *in vivo* and *in vitro* experimental assays.<sup>640,641</sup> Interested readers are referred to several reviews<sup>532,640,642–645</sup> for a comparative analysis of different methods, descriptors, and other related parameters involved in the development of *in silico* models for predicting HIA absorption and oral bioavailability. Machine learning methods have also been widely applied in this area.<sup>646–648</sup> In general, in the vast majority of published studies, the prediction accuracy for HIA is within an acceptable range, with the largest outstanding issue being the predisposition to overfitting induced by the limited data available. Several techniques have been applied to alleviate the small sample size problem and achieve a more reliable performance. For example, Basant et al.<sup>649</sup> recently reported the use of ensemble learning techniques to obtain a reliable model for predicting the HIA of diverse chemicals. Two ensemble learning techniques were utilized: bagged and gradient-boosted trees (see section 2.1.2.4 for details). The proposed ensemble learning-based prediction models were evaluated under several stringent external validation condition and displayed an excellent ability to produce robust HIA models. Using an alternate approach, Shin et al.<sup>650</sup> recently presented a DNN-based *in silico* model that predicted the absorption potential of chemical compounds from molecular descriptors, showing a modest improvement over existing methods without the need for explicit feature engineering.

An alternative method for addressing the problem of overfitting is to use a large and diverse initial data set. A recent study by Wang et al.<sup>651</sup> may serve as a useful starting point for the prediction of HIA for drug candidates in practical applications. A relatively large and structurally diverse Caco-2 cell permeability data set consisting of 1272 molecules was carefully collected under several filtering rules. The utility of this data set has been confirmed by the authors through the production of robust and reliable prediction models.

**6.2.2. Distribution.** Distribution is the process by which drug molecules diffuse or are actively transported from the

bloodstream to body tissues, particularly the tissue(s) where the action is expected to occur. The same features affecting drug absorption (e.g., lipophilicity, molecular size, pH, ionization, etc.) also influence the rate and extent to which the drug distributes to various tissues.

Other factors play a role, particularly nonspecific and reversible protein binding to plasma proteins, including human serum albumin, lipoprotein,  $\alpha$ -acid glycoprotein, and others.<sup>652–655</sup> Plasma protein binding (PPB) influences the volume of distribution, pharmacology, and pharmacokinetics of the drug, as only unbound drug molecules can reach their biological target and exert the intended therapeutic effects.<sup>656–658</sup> Therefore, the fraction of chemical unbound by plasma proteins ( $F_{ub}$ ) is a key parameter in distribution modeling. Previously, our group<sup>659</sup> and other researchers<sup>660–663</sup> have developed models to predict human PPB using machine learning methods such as ANNs, RFs, and SVMs.

Recently, Basant et al.<sup>664</sup> have developed ensemble machine learning models for a four-category classification and prediction of the plasma protein binding affinity of diverse compounds, achieving a classification accuracy of 93% and RMSE of 0.92 for a large PPB data set consisting of 930 compounds. Sun et al.<sup>665</sup> used a heterogeneous training set comprising of 967 diverse chemicals with plasma protein-bound fraction ( $F_b$ ) to build prediction models with six machine learning methods (RF, boosting tree, multiple linear regression, kNN, support vector regression, and ANN) and 26 physicochemical and structural descriptors, obtaining good performance for the whole test set with mean absolute error (MAE) values ranging from 0.126 to 0.178. Ingle et al.<sup>665</sup> explored the advantages of using available pharmaceutical data and machine learning algorithms to provide reliable  $F_{ub}$  prediction. The authors determined that these data improved model performance for pharmaceutical compounds and that model performance for compounds displaying high protein binding affinity was particularly good.

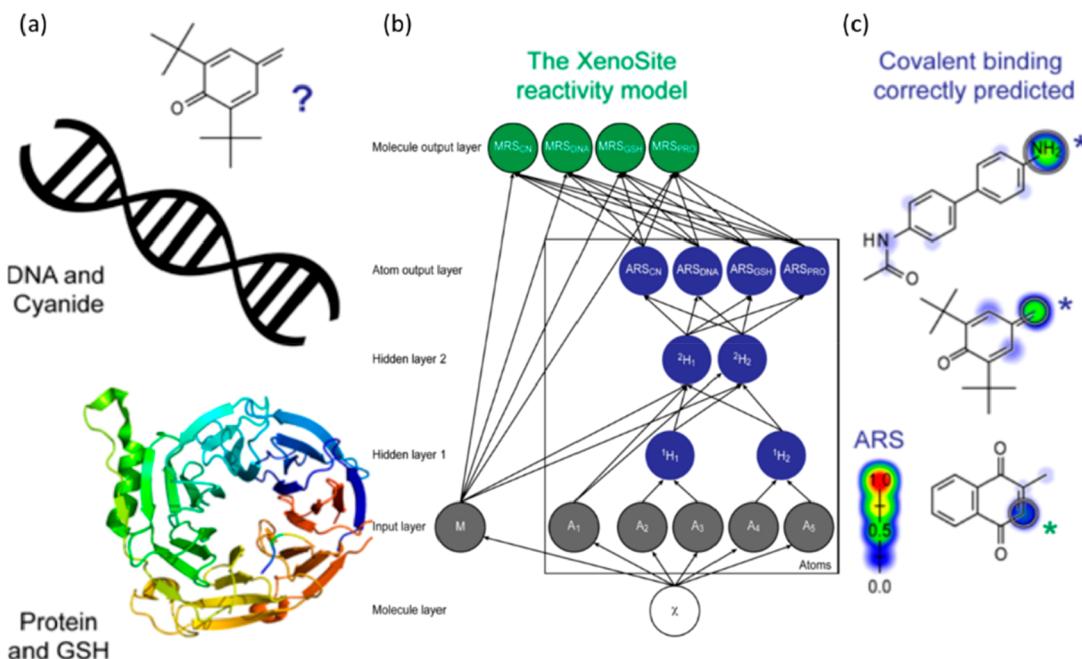
Although substantial effort has been expended to generate experimental data on PPB, obtaining complete, uniform, and consistent coverage of medicinal chemical space is problematic due to heterogeneity in the intent of studies that produced relevant data sources. Sun et al.<sup>663</sup> conducted a thorough comparison and compilation of data sets from three major PPB data sources, namely, the study by Votano et al.<sup>666</sup> (Votano), the database of pharmacokinetic properties (PKDB),<sup>667</sup> and Drugbank.<sup>668</sup> Consequently, a large, heterogeneous data set was acquired and curated for widespread use. It is publicly available and contains 1209 compounds with  $F_b$  values from the three aforementioned sources (Votano, PKDB, and DrugBank). With this extensively curated data set, the construction of regression models for predicting plasma protein-bound fractions seems possible.

Of the various factors that determine the drug distribution, “physical barrier” tissues are often investigated (e.g., the blood-brain-barrier and placental barrier). The blood-brain barrier plays key roles in protecting the brain from changes in blood composition (e.g., hormones, neurotransmitters, and other xenobiotics) and in brain homeostasis. Inconveniently, it also creates a major obstacle to the success of drug development efforts for the treatment of central nervous system disorders. The rate of transfer of a drug molecule into the brain from the blood is mainly determined by its membrane permeability. In recent years, machine learning methods have been utilized to

minimize the number of high-cost, time-consuming, and experimentally challenging blood-brain barrier permeation experiments required in development efforts.<sup>669–671</sup> Although a large number of in vivo and in vitro parameters are available for measuring blood-brain barrier penetration, the most popular and well-regarded parameter remains the ratio of concentrations of the drug in the brain to the concentration in the blood, where a higher ratio corresponds to a higher relative concentration in the brain.<sup>672</sup> Several descriptors have been shown to correlate with blood-brain barrier permeation and have been widely used to construct prediction models, including physicochemical properties (e.g., lipophilicity and ionization state), charge descriptors, and geometrical/topological features (e.g., the number of hydrogen bond donors and acceptors, polar surface area, and number of rotatable bonds).<sup>633,673–677</sup> The establishment of a reliable relationship between molecular descriptors and measured blood-brain barrier permeation remains challenging, as existing publicly available data sets that would allow its estimation are somewhat lacking, both in the quantity and consistency of data.<sup>678</sup> Moreover, our understanding of the complex mechanisms limiting permeation of the blood-brain barrier is still relatively limited compared to our knowledge of intestinal permeability, for instance.<sup>679</sup> The permeation mechanisms in eukaryotic systems mainly consist of passive diffusion, efflux transport, and active transport.<sup>535</sup> Passive diffusion is regarded as the most common mechanism of drug passage through the blood-brain barrier,<sup>673</sup> and therefore, most of the permeation prediction models built to date have focused on this mechanism, to the exclusion of others.<sup>674,680</sup>

Considering the multiple pathways by which drugs cross the blood-brain barrier and the difficulty in predicting permeability from physicochemical features of molecules, Gao et al.<sup>681</sup> recently introduced a novel approach for the prediction of drug blood-brain barrier penetration from drug clinical phenotypes. The central nervous system side-effects of drugs are closely associated with blood-brain barrier permeation. The construction of predictive models using heterogeneous data may be useful to improve performance, based on the relative abundance of side-effect data available in the literature as a result of pharmacovigilance efforts.<sup>682</sup>

**6.2.3. Metabolism.** The chemical transformation of xenobiotics in the liver (as well as other tissues and organs), i.e., metabolism, is usually an enzyme-mediated process, and, considering the versatility of metabolic enzymes, the diversity of the substrates and inhibitors, and the complexity of metabolic processes, has proven challenging to predict. Metabolism determines the fate of a drug in the body and consequently influences its safety profile and therapeutic effect. It plays key roles in many processes, such as the loss of efficacy of many drugs, idiosyncratic adverse drug interactions, hepatotoxicity, and drug–drug interactions.<sup>683,684</sup> Cytochrome P450 isoforms (CYP450s) and UDP-glucuronosyltransferases (UGT), which contribute to phase I and II metabolism, respectively, metabolize the majority of drug molecules.<sup>685–687</sup> CYP450 metabolism accounts for the primary metabolism of approximately 58% of drugs approved by the FDA from 2011 to 2015, an increase from 43% in 2006–2010.<sup>688</sup> Hence, CYP450 inhibition assays are routinely performed to detect those molecules that may have undesired pharmacokinetic properties during the early stage of drug discovery. In recent years, interest in screening metabolism mediated by non-CYP450 phase I metabolic enzymes [e.g., alcohol dehydrogen-



**Figure 33.** (a) A diverse database containing known molecules reacting with biological macromolecules (protein and DNA) and simple nucleophilic traps (cyanide and glutathione (GSH)) was compiled for model training to predict the likelihood that a particular molecule will react with these four substrates. (b) The structure of the XenoSite reactivity model, which is based on a deep CNN architecture. (c) Examples of two types of sites of electrophilic reactivity. From top to bottom, DNA conjugation of *N*-acetylbenzidine, protein conjugation of a butylated hydroxytoluene metabolite, and a glutathione conjugation of menadione are shown. These predictions range from zero to one, indicating the probability that an atom reacts with each of the four nucleophilic targets. Reproduced from ref 709. Copyright 2016 American Chemical Society.

ase (ADH)] and non-UGT-mediated phase II enzymes [e.g., sulfotransferases (STs), *N*-acetyltransferase-1 (NAT1), *N*-acetyltransferase-2 (NAT2), and glutathione S-transferase (GST)] has increased.<sup>689,690</sup>

Research groups have employed various computational approaches (protein–ligand docking, molecular dynamics, molecular mechanics-generalized Born surface area continuum solvation, quantum mechanics/molecular mechanics, etc.) on different targets (CYP450s, UGTs, SULT, and ADH) to study different aspects of the metabolic process (sites/products of metabolism and inhibition/induction of metabolic enzymes). Several reviews have covered aspects of drug metabolism prediction.<sup>684,691–693</sup> Using these approaches, researchers have investigated the details of metabolic mechanisms for a particular system by explicitly describing the electrons involved in breaking and forming molecular bonds. However, to make predictions for a large number of molecules according to their likelihood of metabolism with a particular enzyme panel, we turn to the application of AI methods.<sup>694</sup>

In a broad sense, the applications of AI in the field of predicting metabolism are divided into three main categories: (1) predicting the location of sites of metabolism, (2) predicting specific isoforms responsible for metabolism, and (3) predicting metabolic pharmacokinetics.

The prediction of sites of metabolism is crucial for a xenobiotic, since it provides key information for the derivation of possible metabolites and, consequently, strategies for the development of safer and more efficacious drugs. Currently, several popular AI-based models for predicting sites of metabolism have been developed and display satisfactory accuracy. For example, RS-predictor<sup>695–697</sup> attempts to identify potential sites of CYP450-mediated metabolism of druglike molecules using a combination of hierarchical

descriptors (393 quantum chemical and atom-based descriptors, and 148 topological descriptors) with a multiple instance learning method that is similar in form to the SVM but with a tailored per-instance error contribution during training. A subsequent investigation<sup>696</sup> has shown that the prediction performance is further improved by combining RS-Predictor and SMARTCyp, a method for predicting sites of metabolism that depends on activation energy estimates from a group of model fragments calculated with high-fidelity density functional theory calculations.<sup>698</sup> Finkelmann et al.<sup>699</sup> utilized a similar approach for in collaboration with Bayer and the Schneider group to develop fingerprint methods and utilize them in an RF-based prediction scheme to predict a per-atom pseudolikelihood of involvement in metabolism, with promising retrospective results.<sup>699</sup>

XenoSite<sup>700,701</sup> is a tool that predicts CYP-mediated sites of metabolism based on ANN and multiple descriptions (topological, quantum chemical, and SMARTCyp). A rapid, effective AI-based predictor, FAME, was developed by Kirchmair et al.<sup>702</sup> to predict sites of metabolism. On the basis of the original FAME predictor,<sup>702</sup> Šicho et al.<sup>703</sup> recently proposed a revised methodology for predicting sites of metabolism and the best models have been integrated into a newly developed software package (FAME 2) which is also freely available from the authors.

The Swamidass group has made progress in predicting sites of metabolism,<sup>700,701,704–706</sup> by modeling the formation of reactive metabolites, e.g., epoxidation, with a deep learning network<sup>707</sup> and estimating the reactivity of diverse chemicals.<sup>708,709</sup> For example, a deep multitask network (see section 2.1.1.7), the XenoSite reactivity model, has been designed by this group<sup>709</sup> to investigate the sites and probability of covalent binding of small molecules to protein and DNA (Figure 33).

Their reactivity model outperformed reactivity indices derived from quantum simulations in both situations. In addition to these studies, the same group has recently reported the first publication utilizing deep learning methods to build a quinone formation model and to predict the formation of quinone metabolites through metabolic oxidation mediated by CYP450 and peroxidases.<sup>710</sup>

In addition to the prediction of sites of metabolism, one would also like to infer which specific isoform(s) of phases I or II enzymes might be responsible for the metabolism of a particular compound. AI methods have been applied to classify the inhibitory profiles of molecules of interest against metabolic enzymes. We reviewed some applications in section 3. Interested readers are also referred to several comprehensive reviews related to the prediction of isoform specificity.<sup>690,711,712</sup> Overall, the prediction of isoform specificity is relatively well-established for CYP450-mediated metabolic reactions<sup>711,713</sup> and several methods and software packages are available for users, including CypRules,<sup>714</sup> MetaSite,<sup>715,716</sup> MetaPred,<sup>717</sup> SMARTCyp (in its modified version),<sup>718</sup> and WhichCyp.<sup>719</sup> Other metabolic enzymes are, however, still somewhat underserved.

Overall, the demand for rapid and effective drug metabolism prediction models has continued to grow. Methodologies for predicting sites of metabolism and specific isoforms responsible for the CYP450-mediated metabolism have improved substantially. Meanwhile, researchers have expressed increasing interest in predicting metabolism mediated by non-CYP450 enzymes, although relatively few publications have reported AI-based predictive models for the site of metabolism, isoform specificity, and kinetic parameters for these enzymes to date. Most of the currently available models/packages (including those for CYP450) make metabolism predictions only for the initial molecules provided. However, many primary metabolites undergo another round of phase I or phase II metabolism to form secondary toxic metabolites (e.g., epoxides formed by the oxidation of polycyclic aromatic hydrocarbons in phase I). Therefore, quantum and molecular mechanics methods have been used to mimic the thermodynamics and kinetics of these reactions, as well as the detailed mechanism involved in the metabolic biotransformation. Quantum chemistry has proven its utility in generating informative molecular representations, achieving sustained accuracy for retrospective predictions on drug optimization campaigns. In order to more realistically predict metabolites and identify problematic compounds in the early stages of drug development, it would be useful to integrate current metabolomics and toxicity information with molecular knowledge of chemical and enzymatic reaction mechanisms to provide a unified, chemically aware metabolic platform.

**6.2.4. Excretion.** Drug excretion is the process of removing drug molecules from the body, which directly affects the plasma concentrations of drugs and their metabolites, and plays an important role in the design of drug regimens. Excretion is a complex process that involves several elimination pathways, such as biliary and renal excretion, each of which is composed of a number of different processes. For example, renal excretion is determined by glomerular filtration, active secretion, and reabsorption processes. The complexity of drug excretion processes has hindered the development of AI-based excretion predictors to date. Recently, using a training set comprising 1096 compounds, Lombardo et al.<sup>720</sup> applied linear (partial least-squares) and nonlinear (RF) algorithms to

predict the volume of distribution in humans. A group of molecular descriptors were carefully selected, and the performance of predictors was assessed using a leave-class-out approach and an independent test set. Berellini et al.<sup>721</sup> used physicochemical descriptors and structural fragments to construct a linear partial least-squares predictor for predicting human plasma clearance, based on data from 754 compounds. The predictor yielded a geometric mean fold error of 2.1, and the predicted compound percentages were 59% and 80% for 2- and 3-fold errors, respectively. Kusama et al.<sup>722</sup> established SVM-based predictors with a set of descriptors calculated from the chemical structures of drugs to predict the major clearance pathways of 141 approved drugs in humans and showed high predictive performance.

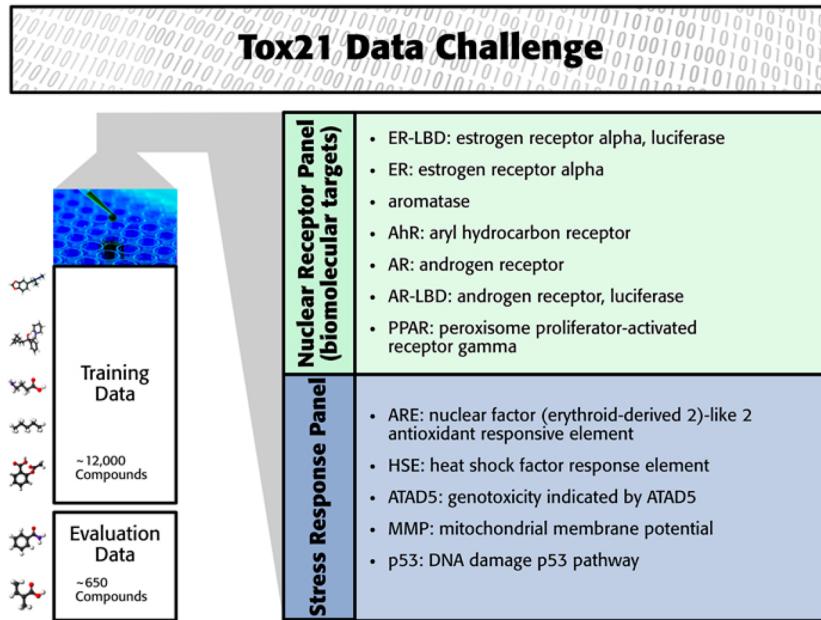
**6.2.5. Toxicity.** Drug toxicity refers to the adverse effect on an organism or a substructure of the organism (e.g., cells and organs) due to the action or metabolism of a compound. Measurements of toxicity are one of the most important and challenging steps involved in the drug discovery and development cycle. Reliable, high-throughput assays are expensive; therefore, computational models that provide a rapid, inexpensive, and reliable alternative to large-scale *in vivo* and *in vitro* bioassays are in high demand. Many web tools and packages have been released to predict toxicity, such as LimTox,<sup>723</sup> pkCSM,<sup>724</sup> admetSAR,<sup>725</sup> Toxtree,<sup>726</sup> and DL-AOT.<sup>727</sup> Meanwhile, advanced AI algorithms have opened up new avenues for predicting toxicity, with quite a few publications discussing the potential roles of AI in this field, such as in predicting toxicological end points or classifying toxic compounds. Most of these algorithms are roughly divided into similarity- and feature-based approaches.

A similarity-based approach predicts the toxicity of a molecule by calculating a matrix of pairwise similarities between compounds, which then is used by predictive approaches such as nearest neighbor algorithms<sup>728,729</sup> and SVM<sup>730</sup> to assign a classification. Similar to most virtual screening efforts, the underlying hypothesis is that similar structures should exert similar biological effects. Feature-based approaches predict toxicity by selecting or weighting input features. Some examples are (generalized) linear,<sup>731,732</sup> RF,<sup>733,734</sup> and naive Bayes models.<sup>735,736</sup>

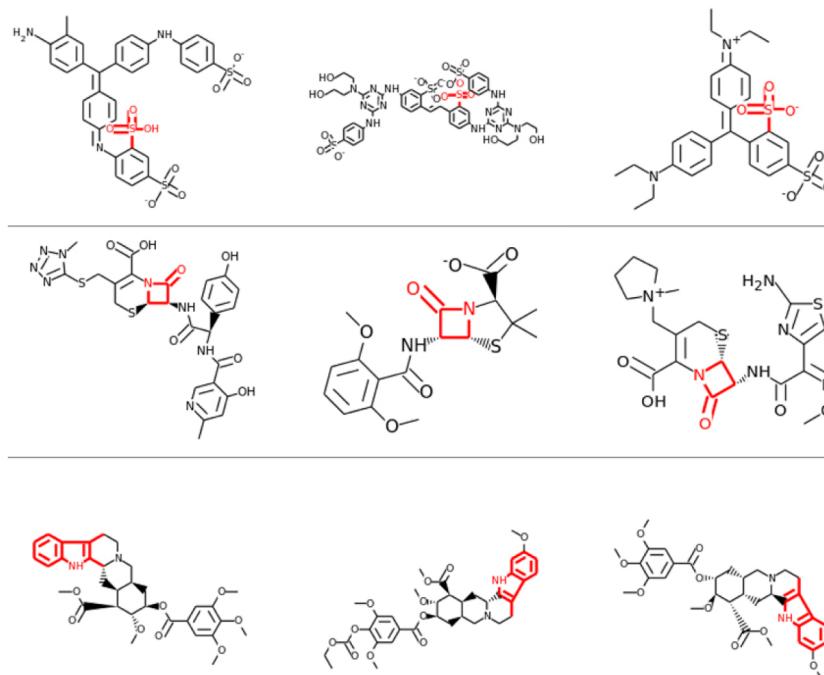
Unlike the feature-based approaches in which the key to a successful prediction is the selection of features that are most relevant to the task, similarity-based approaches require a useful similarity measurement between molecules represented as feature vectors or two-dimensional/three-dimensional graphs. Different approaches exist for identifying structural, physicochemical, and biochemical patterns in these representations, such as toxicophore searching,<sup>737</sup> frequent subgraph mining,<sup>738</sup> and graph kernels.<sup>739</sup> When trained on data from toxicity experiments, AI predictors based on these approaches have been constructed with good overall performance levels.

Many applications exist using either feature- or similarity-based approaches. For example, Pires et al.<sup>724</sup> introduced a novel approach (pkCSM) (Figure 17) utilizing graph-based structural signatures as a molecular representation to predict pharmacokinetic and toxicity properties. Zhang et al.<sup>730</sup> constructed predictive models for drug-induced liver toxicity with five different machine learning methods and MACCS or FP4 fingerprints, and the best model yielded an overall accuracy of 75.0% for the external validation data set.

In recent years, Hou and co-workers<sup>740–742</sup> have reported a series of toxicity evaluations using various machine learning



**Figure 34.** Overview of the Tox21 challenge data set including 12,707 environmental compounds and drugs that were assessed for 12 different toxic effects using specifically designed high-throughput toxicity assays. Reproduced from ref 53. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2016.

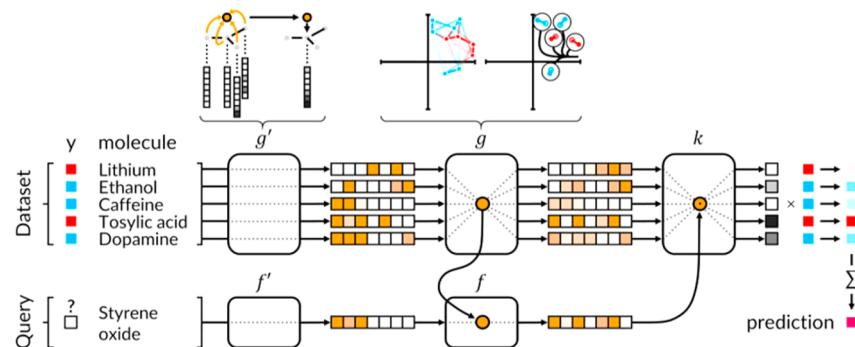


**Figure 35.** Toxicophores learned at different layers of the DeepTox network. Rows 1 and 2 were learned from the first hidden layer, and row 3 was learned from a deeper layer. Reproduced from ref 53. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2016.

methods such as relevance vector machine (RVM),<sup>743</sup> SVM, regularized-RF, C5.0 trees,<sup>744</sup> eXtreme gradient boosting (XGBoost),<sup>745</sup> AdaBoost, SVM boosting (SVMBoost), and RVM boosting (RVMBoost). The toxicity end points include rat oral acute toxicity, respiratory toxicity, and urinary tract toxicity. Considering the overall performance on this set of diverse tasks, the SVM approach is consistently effective for the construction of reliable toxicity prediction models, achieving a 10-fold cross-validation square of the Pearson correlation

coefficients of approximately 0.7 in test sets. In addition, both boosting algorithms generated a strong toxicity predictor from an ensemble of weak predictors. Another ensemble technique was reported by Tharwat et al.<sup>746</sup> in 2016. In this case, four different toxic effects (risk factors) of approved drugs were assessed with a Bagging classifier.

Although prediction models with good performance have been developed using the aforementioned methods, a key limitation is the inability to automatically learn and extract



**Figure 36.** Schematic of the network architecture for one-shot learning in drug discovery. Reproduced from ref 277. Copyright 2017 American Chemical Society.

task-specific chemical features. The development of accurate prediction models critically depends on appropriate molecular representations. In recent years, the emergence of deep learning techniques<sup>53,55,223</sup> has led to novel approaches for molecular representation with the ability to generate new, abstract, and task-specific features. These characteristics make deep learning methods well-suited to predicting toxicity (for a formal description of deep neural networks, see section 2.1.2), as evidenced by their success in the Tox21 Data Challenge.

The Tox21 Data Challenge (Figure 34), which was initiated by United States federal agencies [National Institutes of Health (NIH), Environmental Protection Agency (EPA), and Food and Drug Administration (FDA)], has been the largest effort in the scientific community designed to assess the performance of various computational methods in predicting toxicity and evaluate the potential value of these tools in reducing the number of in vitro and in vivo experiments to date. Mayr et al.<sup>53</sup> participated in this challenge to assess the performance of multitask deep learning methods in predicting toxicity. The authors won this challenge with a prediction pipeline named DeepTox,<sup>53</sup> finding that deep learning supported the automatic learning of features that were similar to well-established toxicophores identified by expert knowledge and experience. As shown in Figure 35, lower layers in the model tended to learn small components of toxicophores, such as sulfonic acid groups (see the highlighted sections in rows 1 and 2 of Figure 35, which were learned in the first hidden layer), whereas in higher layers, these components merged into substructures or whole toxicophores (see row 3 in Figure 35, which was learned from a deeper layer).

Mayr et al.<sup>53</sup> credited their win in the Tox21 challenge to the multitask learning techniques. The authors suggested that multitask learning allows one task to “borrow” features from related tasks and thereby substantially improve the overall performance. This finding is supported as an incidental result of a study by Ramsundar et al. on multitask target prediction.<sup>406</sup> The Pande group<sup>747</sup> also recently reported the robustness and superiority of multitask deep networks over RF methods on four collections of pharmaceutical data (Kaggle, Factors, Kinase, and UV data set collections).

In most deep learning applications, including those described above, the encoding of a particular molecule is represented by a fixed-length vector, and thus the success of deep learning methods strongly depends on good encoding functions for mapping molecular structural information into vectors.<sup>601</sup> Several groups have proposed and described deep learning architectures that predict toxicity through automati-

cally learning relevant features from raw molecular graphs to overcome this limitation. Typical examples of these architectures include undirected graph recursive neural networks for predicting drug-induced liver injury<sup>748</sup> and the improved molecular graph encoding convolutional neural networks (MGE-CNN) architecture for predicting acute oral toxicity,<sup>727</sup> which are similar to the architectures designed to predict aqueous solubility that are discussed in section 6.1.2.

The success of deep learning in predicting toxicity and in other areas of the drug discovery and development process strongly depends on the quantity and quality of input data. For example, the use of massive multitask networks for virtual screening has been shown to significantly improve the prediction accuracy oversimpler machine learning algorithms, but these networks require training on large data sets containing millions of data points.<sup>406</sup> Focusing on the issue of learning meaningful chemical information from only a few data points, Alata-Tran and co-workers introduced the “one-shot learning” technique to enable improved predictive power for tasks with sparse data.<sup>277</sup> A new deep learning architecture (Figure 36), the iterative refinement LSTM combined with a graph convolutional neural network, was introduced to train these models by transferring information between related, but distinct tasks. Three key components are required for the success of one shot-learning. First, the impact of sparse training data is minimized through the use of a similarity metric to compare new data points to the available data and subsequent property imputation for these new data points. Second, a meaningful distance metric is learned to make this similarity transferrable, such that information is traded between query examples and support set elements. Finally, a flexible and meaningful data representation is used as input, which is obtained with graph-convolutional networks (the graph convolution, the graph pool and the graph gather; for more information, see Figure 3 in ref 277). This architecture substantially increased the performance of models trained on limited subsets of the Tox21 and SIDER collections, thereby recovering information that is normally lost with fewer input data.

Overall, much more research is needed to establish a globally applicable and reliable ADMET predictor.<sup>749</sup> Several factors should be considered when constructing reliable AI predictors, such as the importance of including multispecies toxicity data and known issues with *in silico* to *in vitro* (or *in vivo*) correlations. Although current AI-based ADMET predictors remain insufficiently accurate to replace *in vitro* measurements in biologically relevant systems, they are useful for pushing

medicinal chemistry in the right direction, enabling a reduction in the number of synthesis cycles required to resolve ADMET-related problems. An important precondition for this approach is the establishment of trust in the AI models through the validation and careful definition of their applicability domains.

## 7. APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN DRUG REPURPOSING

Faced with skyrocketing costs for developing new drugs, researchers are investigating methods to use drugs that have been developed for one condition and “reposition” them to tackle another. At a fundamental level, repurposing opportunities exist because many drugs have multiple targets,<sup>750,751</sup> and different diseases may share genetic factors, molecular pathways, and clinical manifestations.<sup>752,753</sup> One motivation behind drug repurposing, also referred to as drug repositioning, reprofiling, therapeutic switching, and drug retasking, is the motivation to address unmet therapeutic needs for rare or neglected diseases.<sup>754</sup> Another motivation is the potential to further market and extend the patent life of a drug, thereby increasing the revenue stream. We refer readers to some recent reviews<sup>755–757</sup> to fully appreciate the rationale and challenges associated with this approach.

Computational drug repurposing has shifted from the development of conventional approaches based on the assessment of chemical similarity<sup>758</sup> and molecular docking<sup>759–761</sup> to the creation of new methods providing insights into the evaluation of drug effects using systems biology approaches.<sup>762</sup> Examples include (1) gene expression and functional genomics-based approaches, such as matching drug indications by disease-specific response profiles based on the Connectivity Map (CMap) gene expression signatures<sup>763</sup> and mRNA expression;<sup>764</sup> (2) phenotype and side-effect-based approaches;<sup>765,766</sup> (3) genetic variation-based approaches, such as the discovery of new potential indications for protein targets from genome-wide association studies (GWAS);<sup>767</sup> and (4) disease-network based approaches,<sup>768</sup> which relate disease information scraped from various public resources to build multilevel networks (e.g., Reactome, KEGG pathways, and text-mining)<sup>769</sup> or through a disease graph constructed from gene expression profiles and protein networks.<sup>770</sup> We refer interested readers to some recent reviews<sup>755–757,771,772</sup> for more details on the various computational methods used in drug repurposing. As chemical and genomic data are rapidly accumulating and becoming increasingly accessible and standardized,<sup>773,774</sup> alongside pharmacological and phenotypic information, drug repurposing is becoming an excellent case study for advocates of the deployment of AI technologies in the pharmaceutical industry. The problem plays to the strengths of AI in terms of capturing informative features from noisy, incomplete, and high-dimensional data. Various AI-based methods have been proposed to identify possible drug repurposing opportunities by integrating diverse information from heterogeneous data sources, examples of which include PREDICT,<sup>775</sup> SLAMS,<sup>776</sup> NetLapRLS,<sup>777</sup> and DTINet.<sup>778</sup>

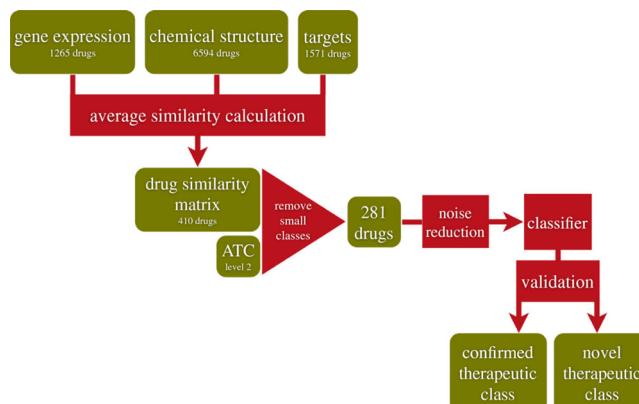
This section focuses on the data-centric AI technologies that are currently used in computational drug repurposing. An exhaustive coverage of the topic exceeds the remit of this section; therefore, instead of a comprehensive description of all publications involved, we classify them into three major categories according to the learning styles used in model construction. Modeling processes and techniques are ad-

dressed using representative examples. We hope this organization will introduce readers to the field of AI-assisted drug repurposing and provide some insights into its promise and potential pitfalls.

### 7.1. Drug Repurposing under the Supervised Learning Paradigm

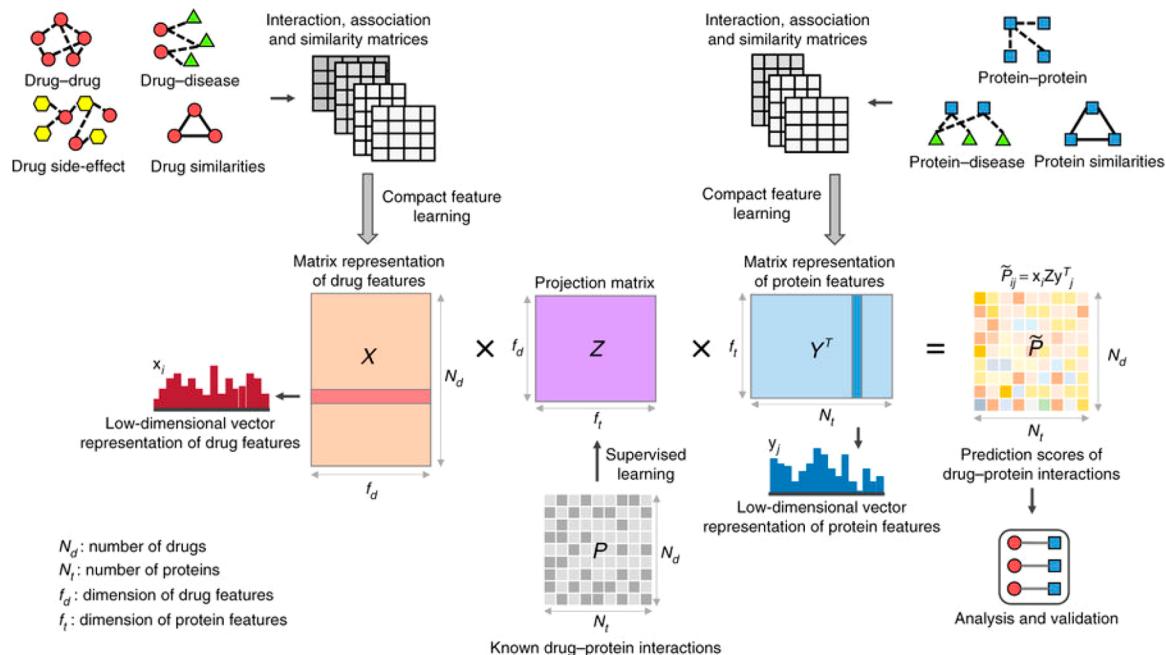
Most AI-assisted drug repurposing studies published to date<sup>776,779–783</sup> have focused on using learning algorithms (e.g., SVM and RF) to develop supervised predictors trained on known associations between drugs, targets, and diseases. These associations are usually represented with various features, including two-dimensional fingerprint similarity, gene expression similarity, and correlation between drug side effects. AI-assisted drugs benefit from large databases of information about small molecules, biological processes, and disease-associated phenotypes.

Although these heterogeneous data sources provide multiple perspectives for predicting novel drug indications, one encounters several difficulties related to data integration, i.e., how to incorporate multiple sources whose relative importance is often unclear. Here, we would like to mention the method proposed by Napolitano et al.,<sup>779</sup> which serves as a good example of overcoming the problem of integrating heterogeneous data. Their approach used three different kernels to integrate different layers of information (Figure 37): (i)



**Figure 37.** Flowchart of the analysis performed by Napolitano et al. The authors constructed three kernels to compute pairwise similarity between compound chemical structures, pairwise compound-similarity based on effects on transcription profiles, and pairwise target similarity based on a statistical analysis of a protein interaction network. These kernels were then integrated to create an overall similarity matrix and analyzed with a SVM classifier to produce Anatomical Therapeutic Chemical classification codes. Green boxes indicate data and red boxes indicate processes. Reproduced with permission from ref 779. (Creative Commons Attribution 2.0 International License: <http://creativecommons.org/licenses/by/2.0/>) Copyright 2013.

structure-based kernel that captures the pairwise similarities in chemical structure and was generated by calculating the distances between the corresponding binary fingerprints; (ii) the transcription-based kernel, which represents drug–drug similarities in terms of their effect on gene expression; and (iii) the target-based kernel, which represents target–target similarities and was built by averaging distances between targets within a protein–protein interaction network. These three kernels were then integrated into a drug similarity matrix that was input into a multiclass SVM classifier. After training to



**Figure 38.** Flowchart of the DTINet pipeline providing and overview of the methods by which it incorporated heterogeneous sources of information. Reproduced with permission from ref 778. (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>) Copyright 2017.

predict Anatomical Therapeutic Chemical (ATC) codes, the SVM classifier finally achieved an average prediction accuracy of 78%. This research provides an efficient method to integrate heterogeneous data sources and reveals the benefits of this approach.

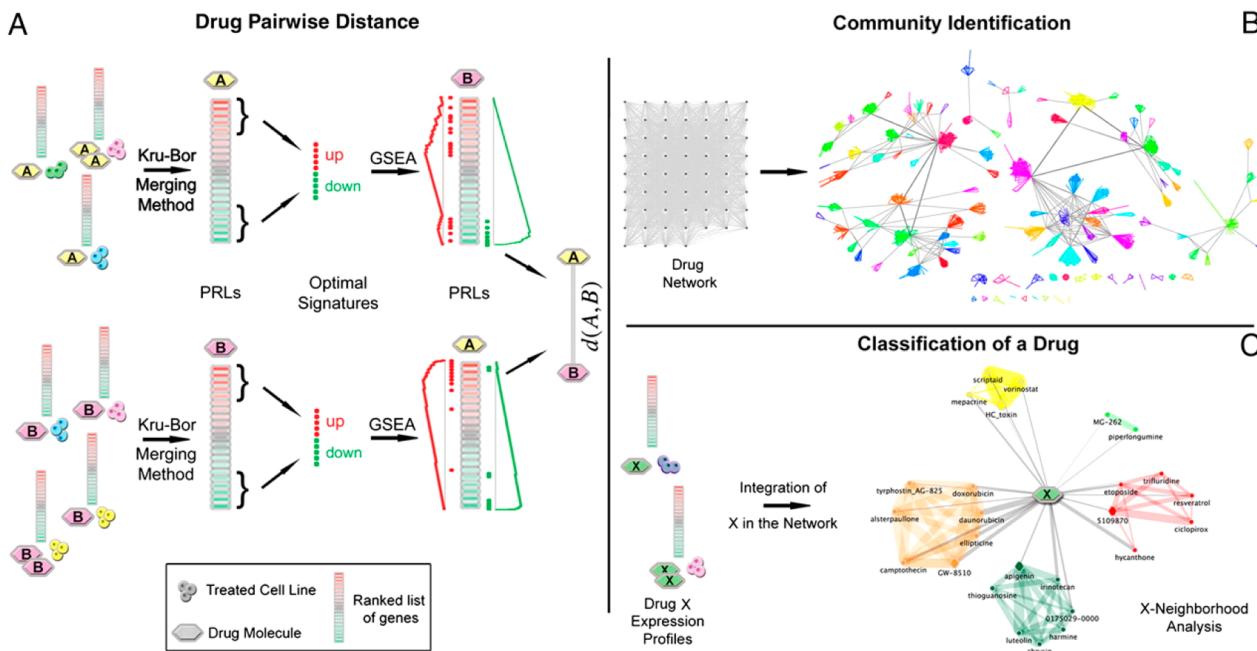
Notably, that AI-assisted drug repurposing models do not necessarily depend on kernel methods, as long as a well-motivated similarity score between a drug and another entity (e.g., drug, disease, or target) under the same model is provided. Various approaches for integration have been developed, and the primary differences are the choice of similarity metrics and how the best method to combine them. For example, a network-based approach has been used for this purpose, where heterogeneous information is fused through a network diffusion process and quantitative or qualitative associations between entities (e.g., gene expression correlation and the presence or absence of an interaction) are directly derived from the obtained diffusion distributions. A good example of this type of network integration approach is the computational pipeline named DTINet proposed by Luo et al.<sup>778</sup> The authors constructed four similarity networks for drugs (Figure 38) based on (i) drug–drug interactions, (ii) drug–disease associations, (iii) drug–side-effect associations, and (iv) chemical structures. Similarly, three similarity networks for proteins were constructed based on (i) protein–protein interactions, (ii) protein–disease associations, and (iii) genome sequences. With these similarity networks, informative but low-dimensional feature vector representations of drugs and proteins were obtained by first applying a network diffusion algorithm (random walk with restart, RWR) separately on individual networks and then jointly optimizing the feature vectors under the compact feature-learning framework. This low-dimensional vector representations obtained from the compact feature-learning framework encoded the relational properties (e.g., similarity), association information, and topological context of each drug (or protein)

node in the heterogeneous network. New drug–target interactions were then predicted based on these representations via a projection from the drug space onto the protein space. By comparing the prediction performance of DTINet with four state-of-the-art methods and verifying those top-ranked predictions experimentally, DTINet has proven to be a useful tool for integrating heterogeneous information to predict new drug–target interactions and repurpose existing drugs.

Supervised models for drug repurposing typically achieve AUC values of 0.75–0.95 in the cross-validation analysis, indicating that they are capable of accurately identifying novel drug–disease associations. Nevertheless, their use in practice, such as for rare diseases, is hindered by the requirement that there are already known and similar drugs for a given condition of interest. Additionally, in some databases, there are no high-quality negative samples available for the training process, leading to a higher false-positive rate than is ideal. Unsupervised and semisupervised learning methods are expected to play an increasingly important role in the field of drug repurposing to meet these challenges.

## 7.2. Drug Repurposing under the Unsupervised Learning Paradigm

Unsupervised computational tools for drug repurposing have several distinct advantages over the supervised tools. They do not require any *a priori* knowledge about drug labels for training data. By discovering hidden patterns in data, these tools can group and identify compounds with similar therapeutic effect. A variety of unsupervised methodologies exist for discovering potential clinical applications of drugs, the most popular of which is the clustering algorithm. This approach was successfully used by Noeske et al.<sup>784</sup> to predict potential activities of known mGluR antagonists. An unsupervised machine learning algorithm (SOM, see section 2.1.2.2) and a topological pharmacophore descriptor (CATS) were used to cluster compounds based on their structure. The



**Figure 39.** Methodological overview of the MANTRA approach. (A) A distance value between each pair of drugs was calculated by a rank-merging procedure and subsequent gene set enrichment analysis. (B) A drug network was generated, where each drug is considered as a node with weighted edges (proportional to distances) connecting pairs of drugs with a distance below a certain threshold. Network communities are identified. (C) For a previously undescribed drug  $X$ , the distance  $d(X, Y)$  is calculated for each drug  $Y$  in the reference data set.  $X$  is added to the network as described in (B). Reproduced with permission from ref 785. Copyright 2010 National Academy of Sciences.

authors identified cross-activities for the mGluR antagonists against other protein targets such as the human dopamine D2-like subfamily, histamine H1, and muscarinic acetylcholine receptors.

Another example that utilized unsupervised techniques for drug repurposing was introduced by Iorio et al.<sup>785</sup> and was later developed into the web tool MANTRA.<sup>786</sup> The drug model in MANTRA was constructed using raw data obtained from the connectivity map (cMap) data set. The core of the approach was to calculate a distance value for each pair of drugs using a rank-aggregation procedure and subsequent gene set enrichment analysis (Figure 39). A drug network was generated in which each node corresponded to a compound, with a weighted edge between two compounds if their corresponding distance was below a preset significance threshold. This network was subsequently mined through a clustering algorithm (affinity propagation algorithm)<sup>787</sup> to identify communities, or modules, of closely related compounds. Drugs within a community share a common mode of action (MoA) and biological pathway. The authors assessed the ability of MANTRA to predict the MoA for anticancer drugs whose gene expression profiles did not appear in the reference CMap data set by quantifying the proximity of the tested drugs to each of the communities. With the use of this network, several cyclin-dependent kinase (CDK) 2 inhibitors were correctly classified and predicted to have a similar MoA to topoisomerase inhibitors. The original method<sup>785</sup> was recently expanded to allow for a supervised refinement of the drug signatures to filter out spurious effects due to secondary effects of the drugs.<sup>788</sup>

Several other unsupervised learning-based methods for drug repurposing have been developed,<sup>789</sup> most of which follow roughly the same paradigm described above. The accuracy of unsupervised methods is usually moderate, due to the limited

use of prior knowledge about known drugs and issues with data heterogeneity. For the same reason, these methods have great prospects in the discovery of new drug-disease associations that are absent from our current understanding of pharmacology, as they provide a relatively bias-free perspective on chemical and biochemical spaces.

### 7.3. Drug Repurposing under the Semisupervised Learning Paradigm

The difficulty in computational drug repurposing lies in the rarity of known associations between entities (drug, target, and disease) and the myriad of unknown associations that remain to be predicted. The number of known associations from public databases is still quite small compared to the entire pharmacological space. In response to this challenge, semisupervised learning<sup>71</sup> has been proposed as a novel approach for drug repurposing due to its reported efficacy in working with small labeled training sets in other disciplines. It tackles this issue by utilizing both the small amount of available labeled information and the abundant unlabeled information to create models with a balance of accuracy and generalizability to new instances.

A widely cited semisupervised learning algorithm for drug repurposing is the Laplacian regularized least-squares (LapRLS).<sup>790</sup> Xia et al.<sup>777</sup> utilized this algorithm to predict drug-target interactions, noting that potential applications of FDA-approved drugs were identified through target prediction. LapRLS constructs two separate spaces for drugs and targets. The drug space is represented by a similarity matrix ( $S_d$ ) obtained from a linear combination of chemical structure similarity and the number of overlapping targets between drug pairs. The target space is represented by a similarity matrix ( $S_t$ ) consisting of sequence similarity and number of common drugs between pairs of targets. LapRLS attempts to estimate the interaction score matrix for drugs ( $F_d$ ) and for targets ( $F_t$ )

based on the drug and target spaces, respectively. For drugs, LapRLS minimizes the squared loss between the known drug-target interaction matrix ( $Y$ ) and  $F_d$  with a regularized term of  $S_d$  and  $F_d$ . This minimization leads to an analytical solution through which  $F_d$  was updated by a rule with  $S_d$  and  $Y$ . The same procedure is performed for targets. Finally, the overall interaction score is obtained by averaging over  $F_d$  and  $F_t$ . The efficiency of the LapRLS approach results from its ability to simultaneously predict the scores of all drug–target pairs. The standard LapRLS was extended by Xia et al. to improve the quality of the predictions<sup>777</sup> by considering drug–target interactions more directly (NetLapRLS). Simply, NetLapRLS and LapRLS use the same regularized least-squares method, except that  $S_d$  in LapRLS is replaced with another matrix considering drug–target interactions.

In addition to the methods mentioned above, other semisupervised learning methods have been used in drug repurposing, such as LPMIHN,<sup>791</sup> NetCBP (network-consistency-based prediction method),<sup>792</sup> and BLM-NII (BLM with neighbor-based interaction-profile inferring).<sup>793</sup> Some encouraging results have been reported, and many methods are available as web-based tools.

However, regardless of whether researchers use supervised, unsupervised, or semisupervised learning approaches, the field of AI-assisted drug repurposing is still in its infancy. In that regard, researchers have not clearly determined which method provides the best results and why that might be the case. The gold standard for these tools, as ever, is their therapeutic viability, i.e., whether insights from *in silico* repurposing efforts suggest candidates with an accuracy that is the same as or better than human experts. A comparison of all of these methods is difficult, due to the varying amounts of data they are able to incorporate and their distinct conceptual foundations. An objective comparison of these methods is a daunting task, since they often use different reference data sets and produce different types of outcomes. The community would benefit from a benchmark data set that covers a wide range of known and confirmed associations, allowing researchers to benchmark model performance, assess predictive power, and conduct error analyses.

## 8. APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN OTHER ASPECTS OF DRUG DISCOVERY

### 8.1. Precision Medicine and Big Data-Driven Drug Discovery

The emergence of large-scale and high-dimensional data incorporating the scientific literature, omics data (genomic, metabolomic, proteomic, and exposomic to name a few), along with physiological and behavioral data obtained by monitoring wearable devices, has opened a route to the simultaneous promotion of drug discovery efforts and personalized medicine. By providing insights into the genetic architecture of a disease in individual patients and the underlying mechanisms of disease pathogenesis, the basic idea of drug discovery and development is shifting from alleviating symptoms to more directly affecting disease progression. Potential drug targets for common and complex diseases can be discovered and confirmed with higher success rates in clinical trials and with lower costs for drug R&D. We refer the interested reader to a recent review on the current major efforts in precision medicine and their potential utility in drug development and drug use.<sup>794</sup>

A thorough and comprehensive search of these massive, heterogeneous, and frequently updated resources is needed to discover possible interactions and leverage “big data” in the era of precision medicine. However, the clarification of ambiguous corresponding “non-genomic” (exposomic and phenotypic) information and development of methods to extract the intrinsic information structure are challenging for curation workflows.

AI methods, particularly deep learning methods, have great potential to push the combination of drug discovery and precision medicine by learning specific intrinsic features of the biomedical big data. To fully harness the potential of AI in guiding the design of new drug molecules, as well as to facilitate tailoring the choice of therapeutic agent to a given patient so as to maximize benefit, research and clinical communities must embrace a common vision of sharing data, with drug-related information made available in a standardized data format in an accessible platform. However, from a methodological and informatics perspective, handling such large volumes of data, particularly unstructured data, remains a substantial challenge,<sup>795,796</sup> although considerable efforts have been made to promote the development of data-sharing technologies and data resources.

In addition, issues surrounding patient privacy and medical ethics, the commercial value of intellectual property, and the currently competitive nature of pharma remain to some extent unresolved. Several programs have been launched by professional associations to generate large data repositories of cancer genomic and clinical data, examples of which are the NCI Genomic Data Commons (GDC) established<sup>797</sup> by NCI Center for Cancer Genomics (CCG) and the AACR Project Genomics Evidence Neoplasia Information Exchange (AACR Project GENIE)<sup>798</sup> established by the American Association for Cancer Research (AACR). Recent efforts to integrate these resources with structural data have resulted in the COSMIC-3D web-service,<sup>799,800</sup> which maps the largest database of somatic mutations in cancer cell lines to their positions in structural models.

With the aid of natural language processing technologies, endeavors are also under way to unlock the value of unstructured or semistructured (for example, scientific papers without an electronic version) data. Several data sources have emerged following the dramatic growth in the use of social media, wellness applications, online tools, wearable devices, and other tools. For example, patient generated health data (PGHD) refers to health data collected by patients (or family members or other caregivers) via wellness and wearable applications. These sorts of meta-data mining operations have been applied widely in epidemiological modeling, producing workable representations of the spread of influenza<sup>801</sup> and contagious illnesses more generally.

Collecting and analyzing large amounts of data to establish actionable patterns is not only viewed as the means of identifying precise treatments in the future but also illuminates a path to more efficient clinical trial protocols.<sup>802</sup> AI-driven protocol design has the potential to make clinical trials more intelligent in a variety of ways (e.g., predicting site performance, preemptively monitoring trial risks, analyzing historical operational data, centralizing trial monitoring, enabling voice-assisted technologies, monitoring drug adherence, enabling site-less and virtual trials, etc.). Furthermore, AI can optimize the overall clinical trial process by helping the industry to more

rapidly recruit patients for clinical trials (particularly trials of rare diseases).

Another notable opportunity for AI technologies is to identify novel pathways, potential therapeutic targets, and drug response biomarkers. As the first step in drug discovery and development, the identification of novel, druggable targets for a therapeutic intervention is a top priority and a formidable challenge for the pharmaceutical and biotech industries, particularly regarding the construction of a robust drug discovery pipeline. Compared to labor, resource, and time-consuming experimental approaches, the use of AI technologies for target identification might exploit genomic “big data” and apply systems-level computational frameworks to drug target discovery, combining gene regulatory information with causal reasoning. Different protocols have been developed to discover targets,<sup>803</sup> a typical example is based on gene expression data from the target tissue, with the intention of identifying cell membrane receptors that play a regulatory role in disease-related gene expression. This approach enables researchers to understand the mechanism of action of a disease and computationally predict the effectiveness of a potential drug target by modulating it. In the last two years, reports of pharmaceutical companies leveraging AI and big data for the purposes of drug target identification have emerged. For example, in June 2017, Genentech announced a collaboration with GNS Healthcare to identify and validate novel cancer drug targets using the company’s proprietary REFS causal machine learning and simulation AI platform.<sup>804</sup> One month later, GSK formed a research partnership with an AI-driven company, Insilico Medicine, to explore how Insilico’s AI capability might help identify novel biological targets and pathways of commercial interest.<sup>805</sup> It is more and more widely accepted that data-driven target identification has the potential to dramatically accelerate the drug development process and bring new treatments to patients much faster.

The identification of biomarkers is an important task not only in medical diagnostics but also in the framework of drug discovery and development programs. With the new trend toward precision medicine, the discovery of novel biomarkers is attracting widespread attention. Currently, several publicly available sources of biological and medical data are available that can be exploited to study biomarkers, examples of which include pharmacogenomics data in preclinical disease models (particularly cell lines) and genomic profiles of clinical samples which have been recently reviewed.<sup>803,806</sup> Like other subdomains in drug discovery, various technology companies and pharmaceutical companies have also joined the field of AI-driven modeling for biomarker development. One example is an agreement, which was announced in October 2017, between Sanofi Pasteur (a global leader in the vaccine industry) and Berg Health (a biopharmaceutical company), under which Berg Health’s “Interrogative Biology” platform and “bAIcis” artificial intelligence tool will be used to identify molecular signatures and potential biomarkers for assessing the immunological response to the Influenza vaccine.<sup>807</sup> Another example of a private-sector initiative applying AI for biomarker development is led by Insilico Medicine. In a recent paper,<sup>808</sup> the company presents a deep learning-based human aging “clock” to assess the biological age of individual patients based on simple hematological measures, with a view of targeting therapeutic interventions.<sup>809</sup> The biomarker model was trained using a large data set of completely anonymized Canadian, South Korean, and Eastern European blood test records. The

company is attempting to identify other biomarkers using deep learning that incorporates blood biochemistry, transcriptomics, and imaging data.

The road to precision medicine is long and winding, although large volumes of biomedical, clinical, and patient data may help to smooth the path somewhat. Precision medicine requires a fresh biological insight into our approach to identifying targets, new therapeutic modalities (e.g., immunotherapy, long-acting RNA interference drugs, CRISPR therapeutics, etc.), molecular imaging technologies, and other developments in related fields. Major efforts<sup>810,811</sup> have resulted in workable drugs for a variety of conditions that were previously neglected, such as rare diseases (e.g., Wolfram syndrome, sickle cell anemia, cystic fibrosis, etc.). An example is the recently approved drug, nusinersen (Spinraza), that is used as a therapeutic oligonucleotide to treat spinal muscular atrophy. Early breakthroughs in precision medicine for oncological conditions have also been achieved, with many targeted cancer therapies (e.g., gene expression modulators and immunotherapies) approved by FDA for the treatment of specific types of cancer, contingent on individual mutations, for example the use of enasidenib in specific cases of acute myeloid leukemia (AML) that contain specific mutations in the isocitrate dehydrogenase II gene.<sup>812</sup>

Autoimmune and inflammatory conditions are also a subject of great interest, due to their strong genetic basis and the availability of techniques to assay resistance signatures. For example, many antigen-specific cell therapies have been investigated for various autoimmune diseases (pemphigus vulgaris, multiple sclerosis, type 1 diabetes, etc.), some of which have entered clinical trials. Another precision medicine approach that is currently in phase two clinical trials is a treatment that blocks the eotaxin-1 protein to prevent inflammation in many autoimmune diseases, including Crohn’s disease and colitis. On the basis of the progress achieved in recent years<sup>813</sup> and the rapid increase in our data acquisition and analysis capabilities, we believe that these approaches will only become more common over time.

## 8.2. Artificial Intelligence-Assisted Molecular Simulations

As described in section 1, computational chemistry is a powerful method for elucidating structural and biochemical characteristics of interest to the modeler in a variety of environments. Molecular dynamics coupled with multiscale quantum mechanics/molecular mechanics methods for monitoring the atomic level time evolution of a biomolecular system have been prominently featured in recent years.<sup>814–816</sup> However, these methods remain too computationally demanding for routine investigations of larger subsets of the chemical space to be explored in current studies. The marriage of AI and computational chemistry offers an alternative with much higher efficiency in principle.<sup>817–819</sup> An established example is the use of Behler–Parinello symmetry function to construct neural network potentials for high-dimensional systems containing thousands of atoms.<sup>820–822</sup> Broader applications have recently been explored, including machine-learned density functional development,<sup>817,823–826</sup> generalized solutions to the Schrödinger equation,<sup>827</sup> predictions of the molecular properties of excited electronic states,<sup>828,829</sup> the classification of chemical trajectory data,<sup>830</sup> many-body expansions,<sup>831</sup> the acceleration of molecular dynamics,<sup>832,833</sup> band gap prediction,<sup>834,835</sup> and high-throughput computational screening for discovering new materials<sup>836–841</sup> and heterogeneous catalysts.<sup>842</sup>

Because of the significant complexity of biosystems and the large number of atoms in biomacromolecules with thousands of degrees of freedom, the applications integrating AI and computational chemistry methods in drug design have encountered several challenges. A remaining issue for AI in bypassing direct energy calculations using quantum mechanics or molecular mechanics methods is how to identify suitable local chemical descriptors that informatively encode atomic environments, such that at the end of training AI predictors provide good predictions on out-of-sample molecules. Another concern arises from the preparation of reference data which needs to be accurate and consistent, ensuring sufficiently low intrinsic errors. First-principles-based methods [e.g., CCSD(T)] are considered able to meet these requirements when handling small organic molecules (e.g., small drug molecules), but these methods are unable to be directly applied to macromolecular targets (e.g., proteins) due to the computational cost; here, best practices remain an open question.

Promising avenues have been proposed, such as introducing a reweighting correction to estimate the results at a desired level of theory with high accuracy (e.g., quantum chemistry methods) based on the results obtained at an inexpensive baseline theory level (e.g., semiempirical quantum chemistry), which has been verified for the prediction of thermochemical properties of molecules<sup>843</sup> and more recently in the estimation of free energy changes in chemical reactions.<sup>844</sup> The appropriate levels of theory for extrapolation are less obvious in biological systems, with an additional error in the poor overlap between sampling spaces at different theory levels. Furthermore, uncertainty about the accurate description of intermolecular specific interactions (e.g., van der Waals forces) between ligands and targets remains, although researchers have long been interested in quantifying the uncertainty with respect to functional choice in energy predictions of protein–ligand interactions.

Despite these challenges, with the continuous development of computational chemistry and AI algorithms, quantum mechanics/molecular mechanics inspired AI predictors will likely be widely used to accelerate the exploration of the chemical space and the identification of novel drug candidates by several orders of magnitude, while maintaining near-quantum mechanical accuracy. This tight integration will ultimately help AI techniques become a truly mainstream tool in pharmaceutical applications.

### 8.3. Data-Driven Explorations of Polypharmacology

The renaissance of AI and the digital revolution in healthcare provide new opportunities for the scientific community to obtain insights into the overall topology and dynamics of disease networks and drug-target and target-disease relationship using massive and multidimensional data (proteomics, clinical, and molecular investigations of patients and disease states, etc.). This approach offers a promising avenue for the identification of new therapeutic targets and for revealing the most relevant targets for a particular disease, helping data-driven polypharmacology discovery become a mainstream paradigm for future drug discovery and development.

Polypharmacology discovery<sup>845,846</sup> refers to the design of a single drug molecule that simultaneously interacts with multiple targets within a disease-related molecular network to achieve the desired therapeutic effects. It is considered a promising method for the design of more effective and less toxic medicines to treat complex diseases (inflammation,

Parkinson's disease, neurodegenerative disease, diabetes, Alzheimer's disease, cancer, etc.). Of course, the accurate construction of polypharmacological profiles is useful for identification of new targets for existing drugs as well, as discussed in the section on drug repurposing (section 7).

In most cases, multiple target combinations are available to control a disease network, and a good combination should be easily modulated by compounds while also achieving a high level of network control. Currently, we do not completely understand the complex interactions between drugs, targets, and diseases in the disease network. However, this type of guilt-by-association network analysis suggests that AI will play a significant role due to its strong data analysis and mining capabilities.

AI might actively participate in addressing another key challenge in polypharmacology discovery: the rational design of multitarget compounds that has historically proven difficult, due to the need to simultaneously optimize the activity of molecules toward several targets, particularly when targets are members of different protein families. One approach is to use AI-assisted virtual screening. A library of virtual compounds is synchronously or sequentially screened to identify drug candidates that might bind to all targets of interest. A well-designed library that covers a wide range of the chemical space with a sufficient number of drug candidates is recommended for this strategy. The ongoing development of better molecular representations will also contribute to the implementation of rational design.

Alternatively, AI-assisted de novo drug design potentially addresses the rational design problem by directly generating molecules with the desired polypharmacological profiles. Compared to virtual screening-based methods that require pre-existing compound libraries, the AI-assisted de novo design usually generates a wider variety of structures, resulting in a higher overall chance of discovering multitarget drugs, particularly in the presence of targets with dissimilar binding pockets and enabling a more focused approach.

Recent advances in multitask and reinforcement approaches in deep learning have raised hopes that AI-assisted de novo polypharmacology discovery may also be conducted in a multiobjective manner, such that one can simultaneously consider both primary (on- and off-target binding of a small molecule to targets of interest) and secondary design constraints (stability, solubility, lipophilicity, synthetic accessibility, ADMET properties, etc.).

One issue is that AI models perform well when each task has sufficient associated data but have proven overconfident in cases where the data are insufficient for an accurate prediction. One way of alleviating the negative impact of data sparsity would be to introduce computational chemistry methods into the model building process. Once the generated molecules bind close to an unexplored region where no structure–property data has been included in the data set, new data points could be produced around that region by selecting a preferred subset from the currently generated structures and then performing, for instance, quantum chemical calculations to create fine-tuned data. Developing an approach which can seamlessly integrate the processes of molecular generation, fine-tuning and adaptive data generation remains an open challenge.

Pharmaceutical companies are aware of the tremendous market potential of AI and big data in polypharmacology discovery. For example, Sanofi signed a \$274 million deal with

Exscientia in 2017 to identify bispecific small molecules to treat diabetes and its comorbidities. Exscientia has also agreed to a research collaboration with Evotec, a German drug company, to discover bispecific cancer immunotherapies. Importantly, while the great potential of artificial intelligence and big data in polypharmacology discovery has caused somewhat of a sensation, the road to success will require considerable further method development.

#### 8.4. Artificial Intelligence-Assisted Evaluation of Synthetic Accessibility

In the process of virtual high-throughput screening or de novo drug design, researchers have used modern tools to quickly create a massive library of virtual compounds with potential biological activity toward a particular biological target. However, the majority of these virtual compounds are either synthetically infeasible or have a developmental synthetic route that scales poorly and, therefore, have limited value and are a high risk for drug developers. It is as much an art as a science for chemists to evaluate synthetic accessibility and determine the most efficient and cost-effective synthetic route for a given synthetically accessible molecule. Chemists must have a solid understanding of organic chemistry and insights obtained through years of lab experience to establish the intuition and expertise required to predict synthetic reactions; nevertheless, planned syntheses often do not result in a drug candidate.<sup>847</sup> Meanwhile, computational chemistry has played an important role in assisting in the prediction of specific chemical reactions but scales poorly to reaction database-sized problems. Providing an automated solution to this goal is not an easy task, and for the past dozen years, numerous groups have painstakingly attempted to achieve this goal with computer algorithms that are built on examples/rules of synthetic reactions and applied to synthetically accessible evaluations, planning of synthetic routes, reaction predictions, and starting material selection.<sup>848</sup> These efforts have produced well-known algorithms such as LHASA,<sup>849</sup> CAMEO,<sup>850</sup> SOPHIA,<sup>851</sup> EROS<sup>852</sup> and SYNCHEM,<sup>853</sup> to name a few. Our group has also introduced a knowledge-based scoring method, named RASA,<sup>854</sup> for the assessment of the synthetic accessibility of drug-like molecules.

Studies from Chematica<sup>855,856</sup> provide a good case-in-point; the team involved spent many years<sup>857</sup> carefully curating data from Reaxys to build a reliable reaction network incorporating 30,000 reaction rules along with information on conditions before beginning large-scale experimental validation, discovering several new one-pot synthetic approaches for known compounds<sup>858</sup> and incorporating pricing information into suggested syntheses.<sup>859</sup> Interested readers may refer to several comprehensive reviews<sup>856,860–863</sup> for a more detailed perspective on rule-based computer-aided organic synthesis.

Despite the tremendous efforts, little evidence is available that these systems have valuable industrial applications. A key limitation is that all these algorithms involve a set of expert rules to predict outcomes. However, it is challenging, even for an expert, to include all reaction rules, since a large organic molecule usually contains multiple reaction sites, and when it reacts with other organic molecules, completely different reactions may occur with respect to the sites involved. As a result, the complexity of chemical reactions has hindered the widespread implementation of these knowledge-based algorithms, and automatic reaction-predictions are rather complicated. Although some degree of automated atom<sup>864</sup> and

reactant-product<sup>865</sup> mapping of mined chemical data is possible, these methods are not yet efficient enough for a “hands-free” analysis.

A data-driven synthetic planning intelligence system might play a useful role in improving the speed of drug discovery due to the high efficiency of AI algorithms in searching a large chemical space. Instead of using handcrafted rules, researchers can build a system that automatically learns how reactions proceed from a large set of examples of synthetic reactions. This system would learn to infer some knowledge of organic chemistry (bond valence, stereochemistry, reaction rules, etc.) to correctly predict the outcome of a reaction. An early model described by Wei et al.<sup>848</sup> used a simple two-step process to predict likely reactions for a particular combination of building blocks, followed by predicting what the outcome of performing that reaction would be. This study was limited to a small number of reactions; later work by Coley et al.<sup>866</sup> adopted a panel of 140,000 reaction templates extracted from the US patent database,<sup>867,868</sup> applied these templates to a pool of reactants to generate plausible products (and augment their data set with “negative” examples), and estimated which candidate product is likely to dominate. They found that their model displays good predictive accuracy, and, importantly, that the accuracy correlates well with the model’s prediction confidence score, indicating that their model is well-calibrated.

Sequence-to-sequence (seq2seq) models have been widely adopted for this task, including the models described by Liu et al.,<sup>812</sup> largely condensing this two-step process into a single step where both processes are implicitly encoded. The model reported by Liu and colleagues was designed to predict reactions by mapping the SMILES strings of reactants into the SMILES string of a product. This prediction is accomplished using an RNN encoder-decoder system with the additive attention model developed by Bahdanau et al. and learning on the US patent database. Their model<sup>258</sup> achieves comparable performance to a rule-based expert system baseline model and alleviates some limitations of such approaches, such as the requirement to conduct explicit atom mapping which is both comprehensive enough to filter inappropriate reactions, and flexible enough to deal with new cases.

Approaches based more directly on chemical logic have also been proposed, as with the work of Segler et al.<sup>869</sup> in using a knowledge-graph approach to teach and model some chemical reasoning and “rediscovery” chemical reactions outside their training data, providing further evidence of the merits of the underlying approach.

The reverse process, by which a retrosynthetic scheme is proposed for a given compound, has also received considerable attention.<sup>870</sup> A retrosynthesis study by Segler et al.<sup>871</sup> has shown that an inverse approach to the method employed by Wei et al., using a somewhat different reaction representation, can successfully be employed to suggest multistep synthetic pathways. Overall, the approach proposed by Segler and colleagues performed considerably better than expert systems in suggesting appropriate pathways and, importantly, in ranking the proposed syntheses sensibly. Further study by the same group<sup>872</sup> combined the Monte Carlo search tree approach<sup>873,874</sup> with three neural networks that define and provide a probabilistic analysis of reaction likelihood and desirability. This approach allows for the formation of a tree in which the root is the final product, the branches are complete synthetic routes, nodes represent a synthetic step, and all precursors required at all prior steps trace back to the root.

The retrosynthesis can be solved if it can be resolved to a purchasable set of building blocks. Monte Carlo tree search enables a random, probabilistic sampling of choices at each step, avoiding extremely expensive enumeration, and allowing the model to learn from previous attempts, where the steps are sampled from machine-learned policies. The authors made use of the proprietary Reaxys database from Elsevier.

As discussed above, most studies published to date have used the US patent database, with a few utilizing proprietary information. Organic reactions reported in the literature are often presented ambiguously, e.g., without full stoichiometry, representing key reagents only as abbreviations, or ignoring small byproducts. As shown in the study by Szymkuć et al.,<sup>856</sup> the number of known reactions for each distinct reaction type is still insufficient to cover all possible combination of substituents, steric and electronic effects, etc. One method to address this limitation is to augment the reported literature with negative reaction examples. Since AI systems need to learn which reactions would or would not occur, the inclusion of negative reaction examples may help achieve higher predictive performance and promote the discovery and refinement of novel chemistry.<sup>866,869</sup>

## 9. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Advances in AI algorithms, particularly in deep-learning approaches, accompanied by increasing architectural hardware specialization (e.g., GPU, TPU, and large-scale parallel computing)<sup>19</sup> and the availability of big data, have catalyzed the rapid development of AI technology, bringing about the third wave of AI.

Achievements in areas such as image and voice recognition, as well as natural language processing, have introduced the field to the wider public and considerably encouraged optimism. In many areas, AI has surpassed human experts in terms of performance. Therefore,<sup>875</sup> its application in drug discovery is not surprising and promising for a field where a conservative approach is often the norm. AI is being integrated into most aspects of the early phase of drug discovery, including target identification, hit/lead discovery, lead optimization, pharmacokinetic property and toxicity prediction, and clinical trial structuring.

Successful examples of deep learning for pharmaceutical problems are coming to the fore. For example, in 2012, Merck sponsored a Kaggle competition to improve the accuracy of molecular property predictions.<sup>52</sup> The winners were not experts in the field, with no particular domain knowledge, and achieved a 15% improvement in relative accuracy over a random forest baseline without explicit feature engineering, using a multitask model. Mayr et al.<sup>53</sup> used a similar architecture to achieve success in the Tox21 challenge, validating the utility of such approaches for various problems in pharmaceutical science. This merging of chemical and computational problems also works in the opposite direction, with much effort expended in the wider AI community to develop tools for the analysis of complex graph representations,<sup>267,876,877</sup> to the benefit of cheminformatics as a discipline.

Despite great initial success, we should keep a clear head for AI's applications in drug discovery, due to the many remaining challenges. Specifically, the acquisition of sufficient, high-quality, problem-specific data remains a major challenge in AI-assisted drug discovery and is a huge barrier to the success

obtained in other fields where AI has been applied. In fields such as machine vision, the data is highly reliable, after all a cat is a cat and not a dog, and researchers are able to compile massive data sets with relative ease as the cost of each data point is small. Unfortunately, this is not the case at all in the drug discovery field, and there are several reasons causing data quality to not be fantastic nor the quantity that large. On one hand, the confidence that can be placed in the labeling of data points highly depends on experimental conditions, which is because biological systems that drugs act in are extremely complicated. Different experimental conditions often give different or even opposite results. On the other hand, the amount of data we have available to us in the drug discovery field is infinitesimal when compared to the vast amounts of information available to, e.g., build vision models or models that relate what you purchased to what else you might want to purchase.

Therefore, the community needs not only a method revolution but also a data revolution in the AI-assisted drug discovery area and thinks about both the data quantity and quality. Imagine if we had one million high quality  $K_d$  values consistently obtained combined with crystal structures, we guess this would come near to solving the docking and scoring problem using AI while singularly advancing human health. However, the question of how we get the data we need at the quality level needed is still an open question. Clearly, better data-sharing is needed, but people do not see an upside in this since open data-sharing arrangements are uncommon in the pharmaceutical sciences, although initiatives such as the Pistoia alliance of large pharmaceutical companies suggests some much-welcomed movement in that regard. Relatively few companies are currently willing to share their data with others, even for negative results or discarded projects. The problems with data quality remain relevant, particularly regarding representation in a unified format. The aforementioned alliance has established the creation of a uniform data format as one of its goals, but it has proven challenging from a technical perspective thus far.

One solution to this dilemma is to develop algorithms that are able to cope with relatively heterogeneous or sparse data sets, of which a few are currently being proposed. Researchers at Stanford University have adapted a type of computer vision algorithm, "one-shot learning", to the problem of drug discovery, resulting in an algorithm that predicts drug properties based on sparse data.<sup>277</sup> Similarly, the transfer learning models developed by the Scheider group at ETH Zurich have been proven to be able to cope with sparse data.

Another significant challenge is to improve the utility and fidelity of molecular representations in order to circumvent problems with the famous 'garbage-in, garbage-out' principle. Most recently developed AI algorithms automatically learn task-related features from "raw" input representations, allowing the internal molecular representation to be refined to an arbitrary standard, and incorporate task specificity as standard. As computations dealing with more complex properties become feasible in a reasonable time frame, at the million-molecule scale, they will likely become increasingly important in model building. Further consideration of the stereochemical configuration, conformational flexibility, surface shape, and overall compactness of the representation is warranted.

As a result of substantial investments of time and capital, and concomitant technological progress, the challenges mentioned above are likely surmountable in the near future. Therefore, a

number of aspects that have impeded the rationalization of drug discovery efforts for the better part of 50 years should achieve more progress in the near future. The opportunities for AI in drug repurposing are particularly fruitful. In the face of the high cost and high risk of failure for R&D efforts with original, chemically innovative drugs, the development of new indications for existing drugs will likely improve the risk-reward profile and offer treatments for historically neglected therapeutic groups. Traditionally, drug repurposing has relied on an empirical approach driven by clinical observations. Currently, large volumes of scientific literature, patents, clinical trial results, and postmarketing authorization data are available for academic and industrial cheminformaticians to leverage. This resource shows promise in the development of a radical, rigorous approach for recontextualizing all of these separate data streams into a more continuous source of revenue and pharmacological insights.

The next promising subdomain is in deep learning-driven virtual screening. Existing computational high-throughput methods suffer from the same problem as their experimental counterparts, a high false-positive rate, due to the imbalance in positive and negative data. Deep learning will be able to make full use of information from receptors, ligands, and their known interactions to aid in the sharing of information from many experiments and multiple targets to improve our per-target performance. In the next few years, we expect to see the development of virtual screening technology using deep learning that replaces or enhances the traditional high-throughput screening process and improves the speed and success rate of screening.

Lead optimization is currently one of the most difficult and complex steps in early phase drug discovery and has traditionally relied on the skills and judgment of medicinal chemists. The fundamental issues here are our grasp of what makes a “good” drug, how to promote good ADMET properties, and how to simultaneously optimize these properties and good on-target activity, while recognizing that these factors are interdependent and sometimes mutually incompatible. With the help of AI, we might ideally perform the simultaneous optimization of all of these parameters and further improve our QSAR models, allowing the rapid selection of the safest compounds for synthesis. Active research in supporting and automating synthesis is ongoing.

Recruiting a sufficient number of suitable patients for clinical trials has long been an issue for pharmaceutical companies. Currently, many clinical trials must greatly extend their time frame at a massive financial cost, as the recruitment of a sufficient number of patients who meet trial criteria is difficult. Using AI to study disease data, pharmaceutical companies might be able to more accurately identify and quickly recruit target patients, depending on the regulatory environment. In addition, AI might also play a role in the management of clinical trials and collection of data from patients, for example, in wearable equipment with an integrated or associated machine learning analysis to improve patient participation, data quality, and operational efficiency in clinical trials.

Applications and publications in this field are currently mushrooming, and various models trained with different machine learning methods and/or different data sets have been reported. There is a tendency to claim the superiority of the latest (usually one’s own) machine learning models, which may leave nonexperts who wish to solve a problem puzzled and helpless when it comes to choosing the best approach. For

example, when studying the toxicity of a molecule of interest, one may be tempted to simply use the model that won the Tox21 competition.<sup>53</sup> However, one of the seemingly less performant models might actually be better suited to predict the toxicity of the compound being considered. Consequently, it will be beneficial for users and developers alike to agree on (and adhere to) objective evaluation standards with independent data sets to estimate the models’ applicability domains and assess their performance when presented with unseen data points.<sup>878,879</sup>

In summary, AI shows potential in multiple fields of drug discovery. As with all concepts, it is unlikely to serve as a panacea, but its adoption should still be increased to assist scientists in their various roles and specialties across the drug development and delivery process. The applications of domain-specific AI in industry are just beginning. We are unlikely to see seismic changes overnight; drug discovery remains a slow business that is predicated on the careful management of risk and on the development of novel science within the constraints of responsibility to the patient and to shareholders. However, the integration of these approaches offers the potential to substantially increase efficiency in some parts of the pipeline, thereby offering researchers time to focus on different problems by offloading simple tasks to combinations of AI and robotics. Furthermore, AI offers potentially useful insights to seasoned scientists based on its extensive “working memory”, which is essentially a novel perspective. Nevertheless, inevitable setbacks and duplications of effort are involved in the process of introducing these changes. We have no doubt, however, that AI will change certain drug discovery processes and likewise that it will advance innovative drug research and development.

## AUTHOR INFORMATION

### Corresponding Authors

\*E-mail: [gisbert@ethz.ch](mailto:gisbert@ethz.ch) (G. Schneider).

\*E-mail: [yangsy@scu.edu.cn](mailto:yangsy@scu.edu.cn) (S. Yang).

### ORCID

Shengyong Yang: [0000-0001-5147-3746](https://orcid.org/0000-0001-5147-3746)

### Notes

The authors declare the following competing financial interest(s): G.S. declares a potential financial conflict of interest in his role as life-science industry consultant and cofounder of inSili.com GmbH, Zurich. No further competing interests are declared.

### Biographies

Xin Yang received her Ph.D. degree (2017) from Sichuan University, where she focused on atomistic simulation under Prof. Ying Xue. After that, she joined the group of Prof. Shengyong Yang at the State Key Laboratory of Biotherapy and Cancer Center, West China Hospital. Her current research interest focuses on integrating machine learning and atomic molecular simulation to problems in protein dynamics, with the goal of leveraging the knowledge gained to design new pharmaceuticals.

Yifei Wang received her B.S. (2018) from the West China School of Pharmacy at Sichuan University, where she performed research on the prediction of drug toxicity based on computational methods. In 2018, she continued her graduate studies under the cosupervision of associate Prof. Linli Li and Prof. Shengyong Yang at Sichuan

University focusing on the application of computational methods to drug design and discovery.

Ryan Byrne received his MSc degree (2015) from Imperial College, London, where he focused on bioinformatics and theoretical systems biology under Profs Michael Sternberg and Michael Stumpf, with an emphasis on genomics and system biology. He briefly worked at the Wellcome Trust Sanger Institute, Cambridge, before beginning studies for his doctorate under the supervision of Prof G. Schneider at the ETH Zurich as a Marie Curie early stage researcher (ESR), focusing on novel representations of small and macro-molecules and on their optimization and utilization for machine-learning tasks.

Gisbert Schneider is a full professor at ETH Zurich, holding the Chair for Computer-Assisted Drug Design. His research focuses on the integration of artificial intelligence into practical medicinal chemistry. His career has led him from the Pharmaceuticals Division at Roche, Basel, to academia, initially to the Goethe-University in Frankfurt where he held the Beilstein Endowed Chair for Chem- and Bioinformatics, and then to his current position at ETH in Zurich. He is an elected Fellow of the University of Tokyo, and the recipient of the 2018 Herman Skolnik Award for his contributions to de novo design of bioactive compounds.

Shengyong Yang received his Ph.D. degree from the School of Chemistry, Sichuan University, in 1999. He was a postdoctoral associate at the Hongkong University of Science and Technology from 1999 to 2001 (with Prof. Zhenyang Lin) and a research scientist at the University of Calgary (Canada) from 2002 to 2005 (with Prof. Tom Ziegler). He is currently a full professor working in the State Key Laboratory of Biotherapy, West China Hospital, Sichuan University. He is an active researcher in the areas of computer-aided drug design (CADD), computational chemistry, medicinal chemistry, and small molecular targeted drug discovery. His current research interests are centered on the development of new methods of CADD and AI and their applications in drug discovery. To date, he has published over 230 peer-reviewed scientific papers.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81573349, 81773633, 81803434, and 21772130), National Science and Technology Major Project (2018ZX09711002-014-002 and 2018ZX09711003-003-006), National Postdoctoral Program for Innovative Talents of China (BX20180205), Post-Doctoral Research and Development Foundation of Sichuan University (2018SCU12040), 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYGDI18001), and Post-Doctor Research Project, West China Hospital, Sichuan University (2018HXBH040). This project has received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Skłodowska-Curie Grant Agreement no. 675555, Accelerated Early staGe drug diScovery (AEGIS).

## ABBREVIATIONS

AAE	adversarial autoencoder
ADMET	absorption, distribution, metabolism, excretion, and toxicity
AI	artificial intelligence
ANN	artificial neural network
Caco-2	human colon adenocarcinoma
CNN	convolutional neural network
DBN	deep belief network

DBM	deep Boltzmann machine
DT	decision tree
DNN	deep neural network
DUD	directory of useful decoys
DUD-E	directory of useful decoys-enhanced
GAN	generative adversarial network
GPU	graphics processing unit
GRU	gated recurrent unit
HIA	human intestinal absorption
HTS	high-throughput screening
kNN	<i>k</i> -nearest neighbor
LapRLS	Laplacian regularized least-squares
LDA	linear discriminant analysis
LSTM	long-short-term memory
LBVS	ligand-based virtual screening
log D	pH-dependent distribution coefficient
log P	octanol–water partition coefficient
log S	aqueous solubility
MC	Monte Carlo
MD	molecular dynamics
MDCK	Madin-Darby canine kidney
MoA	mode of action
PAMPA	parallel artificial membrane permeability assay
PCA	principal component analysis
PPB	plasma protein binding
QSAR	quantitative structure–activity relationship
QSPR	quantitative structure–property relationship
RBF	radial basis function network
RBM	restricted Boltzmann machine
R&D	research and development
RF	random forest
RMSE	root-mean-square error
RNN	recurrent neural network
R <sub>P</sub>	Pearson correlation coefficient
SMILES	simplified molecular input line entry specification
SD	standard deviation
SVM	support vector machine
SBVS	structure-based virtual screening
SOM	self-organizing map
T <sub>m</sub>	melting point
TPU	Google’s tensor processing unit
VAE	variational autoencoder

## REFERENCES

- (1) Nosengo, N. New Tricks for Old Drugs. *Nature* **2016**, *534*, 314–316.
- (2) Kantarjian, H. M.; Prat, F.; Steensma, D. P.; Kurzrock, R.; Stewart, D. J.; Sekeres, M. A.; Leveque, J. Cancer Research in the United States: A Critical Review of Current Status and Proposal for Alternative Models. *Cancer* **2018**, *124*, 2881–2889.
- (3) Moore, T. J.; Zhang, H.; Anderson, G.; Alexander, G. C. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015–2016. *JAMA Inter. Med.* **2018**, *178*, 1451–1457.
- (4) Martin, L.; Hutchens, M.; Hawkins, C.; Radnov, A. How Much Do Clinical Trials Cost? *Nat. Rev. Drug Discovery* **2017**, *16*, 381–382.
- (5) McCarthy, J.; Hayes, P. Some Philosophical Problems From the Standpoint of Artificial Intelligence. In *Machine Intelligence*; Edinburgh University Press: Edinburgh, 1969.
- (6) Barr, A.; Feigenbaum, E. A.; Cohen, P. R. *Handbook of Artificial Intelligence*; Addison-Wesley Longman: Boston, MA, USA, 1982.
- (7) Popovic, D.; Bhatkar, V. P. *Methods and Tools for Applied Artificial Intelligence*; Marcel Dekker: New York, 1994.

- (8) Bobrow, D. G. Natural Language Input for a Computer Problem Solving System. In *Semantic Information Processing*; MIT Press: Cambridge, 1968.
- (9) Weizenbaum, J. ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45.
- (10) Baum, E. B. On the Capabilities of Multilayer Perceptrons. *J. Complex.* **1988**, *4*, 193–215.
- (11) Huang, S.-C.; Huang, Y.-F. Bounds on the Number of Hidden Neurons in Multilayer Perceptrons. *IEEE Trans. Neural. Netw.* **1991**, *2*, 47–55.
- (12) Rumelhart, D.; Hinton, G.; Williams, R. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, 1986.
- (13) Qian, N.; Sejnowski, T. J. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.* **1988**, *202*, 865–884.
- (14) Schwartz, T. Expert Systems Prove Adept at Physics: Recent Successes Demonstrate the Potential to Operate Instrumentation and Distribute Knowledge in the Laboratory. *Comput. Phys.* **1988**, *2*, 40–45.
- (15) Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–38.
- (16) Pearl, J. Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society*, 1985, 329–334.
- (17) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- (18) Quinlan, J. R. Simplifying Decision Trees. *Int. J. Man-Mach. Stud.* **1987**, *27*, 221–234.
- (19) Gaweñ, E.; Hiss, J.; Brown, J.; Schneider, G. Advancing Drug Discovery via GPU-Based Deep Learning. *Expert Opin. Drug Discovery* **2018**, *13*, 579–582.
- (20) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489.
- (21) Moravčík, M.; Schmid, M.; Burch, N.; Lisý, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; Bowling, M. Deepstack: Expert-Level Artificial Intelligence in Heads-up No-Limit Poker. *Science* **2017**, *356*, 508–513.
- (22) Batsford, T. Calculating Optimal Jungling Routes in DOTA2 using neural networks and genetic algorithms. *Game Behaviour* **2014**, *1*, <https://computing.derby.ac.uk/ojs/index.php/gb/article/view/14> (accessed June 20, 2019).
- (23) Hammett, L. P. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
- (24) Hansch, C.; Fujita, T. p- $\sigma$ - $\pi$  Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
- (25) Miller, E.; Hansch, C. Structure-Activity Analysis of Tetrahydrofolate Analogs Using Substituent Constants and Regression Analysis. *J. Pharm. Sci.* **1967**, *56*, 92–97.
- (26) Kopecký, J.; Boček, K.; Vlachova, D. Chemical Structure and Biological Activity on M-and P-Disubstituted Derivatives of Benzene. *Nature* **1965**, *207*, 981–981.
- (27) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726–735.
- (28) Abe, H.; Kumazawa, S.; Taji, T.; Sasaki, S. I. Applications of Computerized Pattern Recognition: A Survey of Correlations Between Pharmacological Activities and Mass Spectra. *Biol. Mass Spectrom.* **1976**, *3*, 151–154.
- (29) Kowalski, B. R.; Bender, C. F. Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test. *J. Am. Chem. Soc.* **1974**, *96*, 916–918.
- (30) Martin, Y. C.; Holland, J. B.; Jarboe, C. H.; Plotnikoff, N. Discriminant Analysis of the Relation Between Physical Properties and the Inhibition of Monoamine Oxidase by Aminotetralins and Aminoindans. *J. Med. Chem.* **1974**, *17*, 409–413.
- (31) Stuper, A. J.; Jurs, P. C. Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques. *J. Am. Chem. Soc.* **1975**, *97*, 182–187.
- (32) Ting, K. L.; Lee, R. C.; Milne, G. W.; Shapiro, M.; Guarino, A. M. Applications of Artificial Intelligence: Relationships Between Mass Spectra and Pharmacological Activity of Drugs. *Science* **1973**, *180*, 417–420.
- (33) Bai, R.; Paull, K. D.; Herald, C. L.; Malspeis, L.; Pettit, G. R.; Hamel, E. Halichondrin B and Homohalichondrin B, Marine Natural Products Binding in the Vinca Domain of Tubulin. Discovery of Tubulin-Based Mechanism of Action by Analysis of Differential Cytotoxicity Data. *J. Biol. Chem.* **1991**, *266*, 15882–15889.
- (34) Paull, K. D.; Lin, C. M.; Malspeis, L.; Hamel, E. Identification of Novel Antimitotic Agents Acting at the Tubulin Level by Computer-Assisted Evaluation of Differential Cytotoxicity Data. *Cancer Res.* **1992**, *52*, 3892–3900.
- (35) Jayaram, H. N.; Gharehbaghi, K.; Jayaram, N. H.; Rieser, J.; Krohn, K.; Paull, K. D. Cytotoxicity of a New IMP Dehydrogenase Inhibitor, Benzamide Riboside, to Human Myelogenous Leukemia K562 Cells. *Biochem. Biophys. Res. Commun.* **1992**, *186*, 1600–1606.
- (36) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386–408.
- (37) Demeler, B.; Zhou, G. Neural Network Optimization for E. Coli Promoter Prediction. *Nucleic Acids Res.* **1991**, *19*, 1593–1599.
- (38) Lukashin, A.; Anshelevich, V.; Amirikyan, B.; Gragerov, A.; Frank-Kamenetskii, M. Neural Network Models for Promoter Recognition. *J. Biomol. Struct. Dyn.* **1989**, *6*, 1123–1133.
- (39) Brunak, S.; Engelbrecht, J.; Knudsen, S. Neural Network Detects Errors in the Assignment of mRNA Splice Sites. *Nucleic Acids Res.* **1990**, *18*, 4797–4801.
- (40) O'Neill; Training, M. C. Back-Propagation Neural Networks to Define and Detect DNA-Binding Sites. *Nucleic Acids Res.* **1991**, *19*, 313–318.
- (41) Holley, L. H.; Karplus, M. Protein Secondary Structure Prediction with a Neural Network. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86*, 152–156.
- (42) Kneller, D.; Cohen, F.; Langridge, R. Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.* **1990**, *214*, 171–182.
- (43) McGregor, M. J.; Flores, T. P.; Sternberg, M. J. Prediction of  $\beta$ -Turns in Proteins Using Neural Networks. *Protein Eng., Des. Sel.* **1989**, *2*, 521–526.
- (44) Weinstein, J. N.; Kohn, K. W.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Monks, A. P.; Scudiero, D. A.; Welch, L.; Koutsoukos, A. D.; Chiausa, A. J.; et al. Neural Computing in Cancer Drug Development: Predicting Mechanism of Action. *Science* **1992**, *258*, 447–451.
- (45) Schneider, G.; Schuchhardt, J.; Wrede, P. Artificial Neural Networks and Simulated Molecular Evolution are Potential Tools for Sequence-Oriented Protein Design. *Bioinformatics* **1994**, *10*, 635–645.
- (46) Schneider, G.; Wrede, P. The Rational Design of Amino Acid Sequences by Artificial Neural Networks and Simulated Molecular Evolution: De Novo Design of an Idealized Leader Peptidase Cleavage Site. *Biophys. J.* **1994**, *66*, 335–344.
- (47) Schneider, G. Generative Models for Artificially-Intelligent Molecular Design. *Mol. Inf.* **2018**, *37*, 1880131.
- (48) Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intelligent Systems and Their Applications* **1998**, *13*, 18–28.

- (49) Ho, T. K. Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, 278–282.
- (50) Lohmann, R.; Schneider, G.; Behrens, D.; Wrede, P. A Neural Network Model for the Prediction of Membrane-Spanning Amino Acid Sequences. *Protein Sci.* **1994**, *3*, 1597–1601.
- (51) Lohmann, R.; Schneider, G.; Wrede, P. Structure Optimization of an Artificial Neural Filter Detecting Membrane-Spanning Amino Acid Sequences. *Biopolymers* **1996**, *38*, 13–29.
- (52) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. *arXiv:1406.1231* **2014**.
- (53) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Frontiers in Environmental Science* **2016**, *3*, 80.
- (54) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.
- (55) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Mol. Inf.* **2016**, *35*, 3–14.
- (56) Zhang, L.; Tan, J.; Han, D.; Zhu, H. From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery. *Drug Discovery Today* **2017**, *22*, 1680–1685.
- (57) Green, C.; Engkvist, O.; Pairaudeau, G. The Convergence of Artificial Intelligence and Chemistry for Improved Drug Discovery. *Future Med. Chem.* **2018**, *10*, 2573–2576.
- (58) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463–477.
- (59) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT press: Cambridge, 2016.
- (60) Murphy, K. P. *Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning*; MIT Press: Cambridge, 2012.
- (61) Bishop, C. M. Periodic Variables. In *Pattern Recognition and Machine Learning*; Springer Science+Business Media: New York, 2006.
- (62) Raymond, J. L.; Medina, J. F. Computational Principles of Supervised Learning in the Cerebellum. *Annu. Rev. Neurosci.* **2018**, *41*, 233–253.
- (63) Hosmer, D. W., Jr.; Lemeshow, S.; Sturdivant, R. X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, 2013.
- (64) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (65) Lewis, D. D. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *European Conference on Machine Learning*, 1998, 4–15.
- (66) Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning*, 1998, 137–142.
- (67) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A k-Means Clustering Algorithm. *J. R. Stat. Soc. C-Apppl.* **1979**, *28*, 100–108.
- (68) Johnson, S. C. Hierarchical Clustering Schemes. *Psychometrika* **1967**, *32*, 241–254.
- (69) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
- (70) Kohonen, T. Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464–1480.
- (71) Chapelle, O.; Schölkopf, B.; Zien, A. *Semi-Supervised Learning. Ser. Adaptive Computation and Machine Learning*; MIT Press: Cambridge, 2006.
- (72) Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, 1995, 189–196.
- (73) Blum, A.; Mitchell, T. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998, 92–100.
- (74) Joachims, T. Transductive Inference for Text Classification Using Support Vector Machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, 200–209.
- (75) Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; Schölkopf, B. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems 6* **2004**, 321–328.
- (76) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discovery Today* **2015**, *20*, 458–465.
- (77) Reker, D.; Schneider, P.; Schneider, G.; Brown, J. Active Learning for Computational Chemogenomics. *Future Med. Chem.* **2017**, *9*, 381–402.
- (78) Chai, H.; Liang, Y.; Wang, S.; Shen, H.-w. A Novel Logistic Regression Model Combining Semi-Supervised Learning and Active Learning for Disease Classification. *Sci. Rep.* **2018**, *8*, 13009.
- (79) Lewis, D. D.; Gale, W. A. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, 3–12.
- (80) Atlas, L. E.; Cohn, D. A.; Ladner, R. E. Training Connectionist Networks with Queries and Selective Sampling. *Advances in Neural Information Processing Systems 2* **1990**, 566–573.
- (81) Cohn, D.; Atlas, L.; Ladner, R. Improving Generalization with Active Learning. *Mach. Learn.* **1994**, *15*, 201–221.
- (82) Hoi, S. C.; Jin, R.; Lyu, M. R. Large-Scale Text Categorization by Batch Mode Active Learning. In *Proceedings of the 15th International Conference on World Wide Web*, 2006, 633–642.
- (83) Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; 1994; Vol. 15, pp 1–67.
- (84) Campbell, C.; Cristianini, N.; Smola, A. Query Learning with Large Margin Classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, 111–118.
- (85) Tong, S.; Koller, D. Support Vector Machine Active Learning with Applications to Text Classification. *J. Mach. Learn. Res.* **2001**, *2*, 45–66.
- (86) Reker, D.; Schneider, P.; Schneider, G. Multi-Objective Active Machine Learning Rapidly Improves Structure–Activity Models and Reveals New Protein–Protein Interaction Inhibitors. *Chem. Sci.* **2016**, *7*, 3919–3927.
- (87) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
- (88) Lang, T.; Flachsenberg, F.; von Luxburg, U.; Rarey, M. Feasibility of Active Machine Learning for Multiclass Compound Classification. *J. Chem. Inf. Model.* **2016**, *56*, 12–20.
- (89) Desai, B.; Dixon, K.; Farrant, E.; Feng, Q.; Gibson, K. R.; van Hoorn, W. P.; Mills, J.; Morgan, T.; Parry, D. M.; Ramjee, M. K.; et al. Rapid Discovery of a Novel Series of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform. *J. Med. Chem.* **2013**, *56*, 3033–3047.
- (90) Fujiwara, Y.; Yamashita, Y.; Osoda, T.; Asogawa, M.; Fukushima, C.; Asao, M.; Shimadzu, H.; Nakao, K.; Shimizu, R. Virtual Screening System for Finding Structurally Diverse Hits by Active Learning. *J. Chem. Inf. Model.* **2008**, *48*, 930–940.
- (91) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
- (92) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. Human-Level Control Through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533.
- (93) Watkins, C. J.; Dayan, P. Q-Learning. *Mach. Learn.* **1992**, *8*, 279–292.
- (94) Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; Silver, D. Rainbow: Combining Improvements in Deep Reinforcement Learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2017.
- (95) Sehnke, F.; Osendorfer, C.; Rückstieß, T.; Graves, A.; Peters, J.; Schmidhuber, J. Parameter-Exploring Policy Gradients. *Neural Netw.* **2010**, *23*, 551–559.
- (96) Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for

- Deep Reinforcement Learning. In *International Conference on Machine Learning*, 2016, 1928–1937.
- (97) Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust Region Policy Optimization. In *International Conference on Machine Learning*, 2015, 1889–1897.
- (98) Moriarty, D. E.; Schultz, A. C.; Grefenstette, J. J. Evolutionary Algorithms for Reinforcement Learning. *Journal of Artificial Intelligence Research* **1999**, *11*, 241–276.
- (99) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
- (100) Torrey, L.; Shavlik, J. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, 2010.
- (101) Bickel, S.; Brückner, M.; Scheffer, T. Discriminative Learning for Differing Training and Test Distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007, 81–88.
- (102) Quattoni, A.; Collins, M.; Darrell, T. Transfer Learning for Image Classification with Sparse Prototype Representations. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, 1–8.
- (103) Wang, C.; Mahadevan, S. Manifold Alignment Using Procrustes Analysis. In *Proceedings of the 25th International Conference on Machine Learning*, 2008, 1120–1127.
- (104) Fei-Fei, L.; Fergus, R.; Perona, P. One-Shot Learning of Object Categories. *IEEE Transactions on Pattern Analysis Machine Intelligence* **2006**, *28*, 594–611.
- (105) Bonilla, E. V.; Chai, K. M.; Williams, C. Multi-Task Gaussian Process Prediction. In *Advances in Neural Information Processing Systems*, 2008, 153–160.
- (106) Stark, M.; Goesele, M.; Schiele, B. A Shape-Based Object Class Model for Knowledge Transfer. In *2009 IEEE 12th International Conference on Computer Vision*, 2009, 373–380.
- (107) Wang, D.; Li, Y.; Lin, Y.; Zhuang, Y. Relational Knowledge Transfer for Zero-Shot Learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, 2145–2151.
- (108) Yao, Y.; Doretto, G. Boosting for Transfer Learning with Multiple Sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, 1855–1862.
- (109) Kovac, K. *Multitask Learning for Bayesian Neural Networks*. Master's Thesis, University of Toronto, 2005.
- (110) Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Mach. Learn.* **1997**, *29*, 131–163.
- (111) Sahami, M. Learning Limited Dependence Bayesian Classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, 335–338.
- (112) Langley, P.; Iba, W.; Thompson, K. An Analysis of Bayesian Classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992, 223–228.
- (113) Domingos, P.; Pazzani, M. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Mach. Learn.* **1997**, *29*, 103–130.
- (114) Zhang, H.; Ling, C. X. Learnability of Augmented Naïve Bayes in Nominal Domains. In *Proceedings of the 18th International Conference on Machine Learning*, 2001, 617–623.
- (115) Rish, I. An Empirical Study of the Naïve Bayes Classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001, 41–46.
- (116) Kononenko, I. Semi-Naïve Bayesian Classifier. In *European Working Session on Learning*, 1991, 206–219.
- (117) Jensen, F. V. *An Introduction to Bayesian Networks*; UCL press: London, 1996.
- (118) Heckerman, D. A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*; Springer: Dordrecht, 1998.
- (119) Pearl, J. Fusion, Propagation, and Structuring in Belief Networks. *Artif. Intell.* **1986**, *29*, 241–288.
- (120) Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.* **2000**, *7*, 601–620.
- (121) Smith, V. A.; Jarvis, E. D.; Hartemink, A. J. Evaluating Functional Network Inference Using Simulations of Complex Biological Systems. *Bioinformatics* **2002**, *18*, S216–S224.
- (122) Yu, J.; Smith, V. A.; Wang, P. P.; Hartemink, A. J.; Jarvis, E. D. Advances to Bayesian Network Inference for Generating Causal Networks from Observational Biological Data. *Bioinformatics* **2004**, *20*, 3594–3603.
- (123) Margaritis, D.; Thrun, S. Bayesian Network Induction via Local Neighborhoods. In *Advances in Neural Information Processing Systems*, 2000, 505–511.
- (124) Garcia, L. D.; Stillman, M.; Sturmels, B. Algebraic Geometry of Bayesian Networks. *J. Symb. Comput.* **2005**, *39*, 331–355.
- (125) Aha, D. W.; Kibler, D.; Albert, M. K. Instance-Based Learning Algorithms. *Mach. Learn.* **1991**, *6*, 37–66.
- (126) Wettschereck, D.; Aha, D. W.; Mohri, T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artif. Intell. Rev.* **1997**, *11*, 273–314.
- (127) Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
- (128) Rupp, M.; Schneider, P.; Schneider, G. Distance Phenomena in High-Dimensional Chemical Descriptor Spaces: Consequences for Similarity-Based Approaches. *J. Comput. Chem.* **2009**, *30*, 2285–2296.
- (129) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422.
- (130) Joshi, A.; Mehta, A. Analysis of k-Nearest Neighbor Technique for Breast Cancer Disease Classification. *Int. J. Recent Sci. Res.* **2018**, *9*, 26126–26130.
- (131) Tang, L.; Pan, H.; Yao, Y. PANK-A Financial Time Series Prediction Model Integrating Principal Component Analysis, Affinity Propagation Clustering and Nested k-Nearest Neighbor Regression. *J. Interdiscip. Math.* **2018**, *21*, 717–728.
- (132) Tang, L.; Pan, H.; Yao, Y. K-Nearest Neighbor Regression with Principal Component Analysis for Financial Time Series Prediction. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, 2018, 127–131.
- (133) Chaudhuri, K.; Dasgupta, S. Rates of Convergence for Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems*, 2014, 3437–3445.
- (134) Berchtold, S.; Böhm, C.; Keim, D. A.; Kriegel, H.-P. A Cost Model for Nearest Neighbor Search in High-Dimensional Data Space. In *the sixteenth ACM SIGACT-SIGMOD-SIGART Symposium*, 1997, 78–86.
- (135) Arya, S.; Mount, D. M.; Narayan, O. Accounting for Boundary Effects in Nearest Neighbor Searching. In *Proceedings of the Eleventh Annual Symposium on Computational Geometry*, 1995, 336–344.
- (136) Thilagaraj, M.; Rajasekaran, M. P. Epileptic Seizure Mining via Novel Empirical Wavelet Feature with J48 and KNN Classifier. In *Intelligent Engineering Informatics*; Springer: Singapore, 2018.
- (137) Xia, C.; Hsu, W.; Lee, M.-L.; Ooi, B. C. BORDER: Efficient Computation of Boundary Points. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 289–303.
- (138) Qiu, B.-Z.; Yue, F.; Shen, J.-Y. BRIM: An Efficient Boundary Points Detecting Algorithm. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2007, 761–768.
- (139) Li, X.; Lv, J.; Yi, Z. An Efficient Representation-Based Method for Boundary Point and Outlier Detection. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *29*, 51–62.
- (140) Chiavazzo, E.; Covino, R.; Coifman, R. R.; Gear, C. W.; Georgiou, A. S.; Hummer, G.; Kevrekidis, I. G. Intrinsic Map Dynamics Exploration for Uncharted Effective Free-Energy Landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5494–E5503.
- (141) MOODY, J. Learning with Localized Receptive Fields. In *Proceedings of Connectionist Models Summer School*; Morgan Kaufmann Publications: San Mateo, CA, 1988.
- (142) Moody, J.; Darken, C. J. Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Comput.* **1989**, *1*, 281–294.

- (143) Powell, M. J. D. Radial Basis Functions Approximations to Polynomials. In *Proceedings of 12th Biennial Numerical Analysis Conference*, 1988, 223–241.
- (144) Powell, M. Radial Basis Functions for Multivariable Interpolation: A Review. In *IMA Conference on Algorithms for the Approximation of Functions and Data*, 1987, 143–167.
- (145) Park, J.; Sandberg, I. W. Universal Approximation Using Radial-Basis-Function Networks. *Neural Comput.* **1991**, *3*, 246–257.
- (146) Chen, S.; Cowan, C. F.; Grant, P. M. Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks. *IEEE Trans. Neural. Netw.* **1991**, *2*, 302–309.
- (147) Wettschereck, D.; Dietterich, T. Improving the Performance of Radial Basis Function Networks by Learning Center Locations. In *Advances in Neural Information Processing Systems*, 1992, 1133–1140.
- (148) Schwenker, F.; Kestler, H.; Palm, G.; Hoher, M. Similarities of LVQ and RBF Learning—A Survey of Learning Rules and the Application to the Classification of Signals from High-Resolution Electrocardiography. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, 1994, 646–651.
- (149) Kubat, M. Decision Trees Van Initialize Radial-Basis Function Networks. *IEEE Trans. Neural. Netw.* **1998**, *9*, 813–821.
- (150) Schwenker, F.; Dietrich, C. Initialisation of Radial Basis Function Networks Using Classification Trees. *Neural Netw. World.* **2000**, *10*, 473–482.
- (151) Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. *Curr. Med. Chem.* **2009**, *16*, 258–266.
- (152) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of De Novo-Designed Chemical Entities Through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 4067–4072.
- (153) Fritzke, B. Growing Grid—A Self-Organizing Network with Constant Neighborhood Range and Adaptation Strength. *Neural Process. Lett.* **1995**, *2*, 9–13.
- (154) Miikkulainen, R. Script Recognition with Hierarchical Feature Maps. In *Connectionist Natural Language Processing*; Springer: Dordrecht, 1992.
- (155) Kohonen, T. Essentials of the self-organizing map. *Neural Netw.* **2013**, *37*, 52–65.
- (156) Kohonen, T.; Kaski, S.; Lagus, K.; Salojarvi, J.; Honkela, J.; Paatero, V.; Saarela, A. Self organization of a massive document collection. *IEEE Trans. Neural. Netw.* **2000**, *11*, 574–585.
- (157) Kohonen, T.; Xing, H. Contextually Self-Organized Maps of Chinese Words. In *International Workshop on Self-Organizing Maps*, 2011, 16–29.
- (158) Domingos, P. A. Few Useful Things to Know About Machine Learning. *Commun. ACM* **2012**, *55*, 78–87.
- (159) Weston, J.; Mukherjee, S.; Chapelle, O.; Pontil, M.; Poggio, T.; Vapnik, V. Feature Selection for SVMs. In *Advances in Neural Information Processing Systems*, 2001, 668–674.
- (160) Tomar, D.; Agarwal, S. A Comparison on Multi-Class Classification Methods Based on Least Squares Twin Support Vector Machine. *Knowledge-Based Syst.* **2015**, *81*, 131–147.
- (161) Fenn, M.; Guerracino, M.; Pi, J.; Pardalos, P. M. Raman Spectroscopy Using a Multiclass Extension of Fisher-Based Feature Selection Support Vector Machines (FFS-SVM) for Characterizing In-Vitro Apoptotic Cell Death Induced by Paclitaxel. In *International Conference on Learning and Intelligent Optimization*, 2014, 306–323.
- (162) Sun, Z.; Guo, Z.; Liu, C.; Wang, X.; Liu, J.; Liu, S. Fast Extended One-Versus-Rest Multi-Label Support Vector Machine Using Approximate Extreme Points. *IEEE Access* **2017**, *5*, 8526–8535.
- (163) Ding, S.; Zhang, X.; An, Y.; Xue, Y. Weighted Linear Loss Multiple Birth Support Vector Machine Based on Information Granulation for Multi-Class Classification. *Pattern Recognit.* **2017**, *67*, 32–46.
- (164) Yang, Z.-X.; Shao, Y.-H.; Zhang, X.-S. Multiple Birth Support Vector Machine for Multi-Class Classification. *Neural Comput. Appl.* **2013**, *22*, 153–161.
- (165) Zhang, X.; Ding, S.; Xue, Y. An Improved Multiple Birth Support Vector Machine for Pattern Classification. *Neurocomputing* **2017**, *225*, 119–128.
- (166) Angulo, C.; Parra, X.; Catala, A. K-SVCR. A Support Vector Machine for Multi-Class Classification. *Neurocomputing* **2003**, *55*, 57–77.
- (167) Tang, L.; Tian, Y.; Pardalos, P. M. A Novel Perspective on Multiclass Classification: Regular Simplex Support Vector Machine. *Inf. Sci.* **2019**, *480*, 324–338.
- (168) Lauer, F.; Guermeur, Y. MSVMpack: A Multi-Class Support Vector Machine Package. *J. Mach. Learn. Res.* **2011**, *12*, 2293–2296.
- (169) Burges, C. J. A. Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121–167.
- (170) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106.
- (171) Quinlan, J. R. Discovering Rules by Induction from Large Collections of Examples. In *Expert Systems in the Micro Electronics Age*; Edinburgh University Press: Edinburgh, 1979.
- (172) Quinlan, J. Learning Efficient Classification Procedures and Application to Chess End Games. In *Machine Learning: An Artificial Intelligence Approach*; Tioga: Palo Alto, CA, 1983.
- (173) Quinlan, J. *C4.5: Programs for Machine Learning*. ; Morgan Kaufmann: San Francisco, 1993.
- (174) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth: Belmont, CA, 1984.
- (175) Kingsford, C.; Salzberg, S. L. What Are Decision Trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013.
- (176) Pedrycz, W.; Sosnowski, Z. A. The Design of Decision Trees in the Framework of Granular Data and Their Application to Software Quality Models. *Fuzzy Sets Syst.* **2001**, *123*, 271–290.
- (177) Podgorelec, V.; Kokol, P.; Stiglic, B.; Rozman, I. Decision Trees: An Overview and Their Use in Medicine. *J. Med. Syst.* **2002**, *26*, 445–463.
- (178) Bradford, J. P.; Kunz, C.; Kohavi, R.; Brunk, C.; Brodley, C. E. Pruning Decision Trees with Misclassification Costs. In *European Conference on Machine Learning*, 1998, 131–136.
- (179) Kotsiantis, S. B.; Zaharakis, I.; Pintelas, P. Supervised Machine Learning: A Review of Classification Techniques. In *Emerging Artificial Intelligence Applications in Computer Engineering*; IOS Press: Netherlands, 2007.
- (180) Caruana, R.; Niculescu-Mizil, A. An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006, 161–168.
- (181) Zhang, C.; Ma, Y. *Ensemble Machine Learning: Methods and Applications*; Springer Science+Business Media: New York, 2012.
- (182) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140.
- (183) Schapire, R. E. The Boosting Approach to Machine Learning: An Overview. In *Nonlinear Estimation and Classification*; Springer: New York, 2003.
- (184) Wolpert, D. H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259.
- (185) Freund, Y.; Schapire, R. E. A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139.
- (186) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232.
- (187) Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378.
- (188) Breiman, L. Stacked Regressions. *Mach. Learn.* **1996**, *24*, 49–64.
- (189) Koren, Yehuda. *Bellkor Solution to the Netflix Grand Prize*. [https://www.netflixprize.com/assets/GrandPrize2009\\_BPC\\_BellKor.pdf](https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf) (accessed June 20, 2019).
- (190) Xu, L.; Jiang, J.-H.; Zhou, Y.-P.; Wu, H.-L.; Shen, G.-L.; Yu, R.-Q. MCCV Stacked Regression for Model Combination and Fast Spectral Interval Selection in Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 226–230.

- (191) Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Computing Surveys (CSUR)*. **2018**, *50*, 94.
- (192) Yan, J.; Zhang, B.; Liu, N.; Yan, S.; Cheng, Q.; Fan, W.; Yang, Q.; Xi, W.; Chen, Z. Effective and Efficient Dimensionality Reduction for Large-scale and Streaming Data Preprocessing. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 320–333.
- (193) Jolie, I. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
- (194) Martínez, A. M.; Kak, A. C. PCA Versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 228–233.
- (195) Fukunaga, K. *Introduction to Statistical Pattern Classification*; Academic Press: USA, 1990.
- (196) Yu, H.; Yang, J. A Direct LDA Algorithm for High-Dimensional Data—with Application to Face Recognition. *Pattern Recognit* **2001**, *34*, 2067–2070.
- (197) Li, H.; Jiang, T.; Zhang, K. Efficient and Robust Feature Extraction by Maximum Margin Criterion. In *Advances in Neural Information Processing Systems*, 2004, 97–104.
- (198) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326.
- (199) Tenenbaum, J. B.; De Silva, V.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323.
- (200) Maaten, L. v. d.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- (201) Wang, X.; Paliwal, K. K. Feature Extraction and Dimensionality Reduction Algorithms and Their Applications in Vowel Recognition. *Pattern Recognit* **2003**, *36*, 2429–2439.
- (202) Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* **1933**, *24*, 417–441.
- (203) Rao, C. R. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya: Ind. J. Stat. A* **1964**, *26*, 329–358.
- (204) Bocchieri, E.; Wilpon, J. Discriminative Analysis for Feature Reduction in Automatic Speech Recognition. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992, 501–504.
- (205) Poston, W. L.; Marchette, D. J. Recursive Dimensionality Reduction Using Fisher's Linear Discriminant. *Pattern Recognit* **1998**, *31*, 881–888.
- (206) Sun, D. X. Feature Dimension Reduction Using Reduced-Rank Maximum Likelihood Estimation for Hidden Markov Models. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, 244–247.
- (207) Van Der Maaten, L. Accelerating t-SNE Using Tree-Based Algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221–3245.
- (208) Burges, C. Geometric Methods for Feature Selection and Dimensional Reduction: A Guided Tour. In *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*; Springer Science+Business Media: New York, 2005.
- (209) Lee, J. A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer: New York, 2007.
- (210) Saul, L. K.; Weinberger, K. Q.; Ham, J. H.; Sha, F.; Lee, D. D. Spectral Methods for Dimensionality Reduction. In *Semisupervised Learning*; MIT Press: Cambridge, 2006.
- (211) Venna, J. *Dimensionality Reduction for Visual Exploration of Similarity Structures*. Doctoral Dissertation, Helsinki University of Technology, 2007.
- (212) Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality Reduction: A Comparative. *J. Mach. Learn. Res.* **2009**, *10*, 66–71.
- (213) Reutlinger, M.; Schneider, G. Nonlinear Dimensionality Reduction and Mapping of Compound Libraries for Drug Discovery. *J. Mol. Graphics Modell.* **2012**, *34*, 108–117.
- (214) Wasserman, P. D. *Neural Computing: Theory and Practice*; Van Nostrand Reinhold: New York, 1989.
- (215) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley/Addison Wesley Longman: Reading, MA, US, 1991.
- (216) Smith, M. Neural Networks for Statistical Modeling. In *Proceedings Of The 19th Annual SAS Users Group International Conference*, 1994, 1538–1550.
- (217) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536.
- (218) Rumelhart, D. E.; Widrow, B.; Lehr, M. A. The Basic Ideas in Neural Networks. *Commun. ACM* **1994**, *37*, 87–93.
- (219) Hinton, G. E. How Neural Networks Learn from Experience. *Sci. Am.* **1992**, *267*, 144–151.
- (220) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* **2015**.
- (221) Werbos, P. J. *The Roots of Backpropagation: from Ordered Derivatives to Neural Networks and Political Forecasting*; John Wiley & Sons: New York, 1994.
- (222) Dreyfus, S. E. Artificial Neural Networks, Back Propagation, and the Kelley-Bryson Gradient Procedure. *J. Guid. Control Dyn.* **1990**, *13*, 926–928.
- (223) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
- (224) LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput* **1989**, *1*, 541–551.
- (225) Goodfellow, I. J.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. Maxout Networks. In *International Conference on Machine Learning*, 2013, 1319–1327.
- (226) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
- (227) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 770–778.
- (228) Meissner, M.; Schmuker, M.; Schneider, G. Optimized Particle Swarm Optimization (OPSO) and Its Application to Artificial Neural Network Training. *BMC Bioinf.* **2006**, *7*, 125.
- (229) Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*; IEEE Press: New York, 2001.
- (230) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **1997**, *9*, 1735–1780.
- (231) Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Workshop on Deep Learning*, 2014.
- (232) Grefenstette, E.; Hermann, K. M.; Suleyman, M.; Blunsom, P. Learning to Transduce with Unbounded Memory. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, 1828–1836.
- (233) Hinton, G. E.; Sejnowski, T. J. Optimal Perceptual Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983, 448–453.
- (234) Smolensky, P. Information Processing in Dynamical Systems: Foundations of Harmony Theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, 1986.
- (235) Hinton, G. E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comput* **2002**, *14*, 1771–1800.
- (236) Salakhutdinov, R.; Hinton, G. Deep Boltzmann Machines. *J. Mach. Learn. Res.* **2009**, *24*, 448–455.
- (237) Hinton, G. E.; Osindero, S.; Teh, Y.-W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput* **2006**, *18*, 1527–1554.
- (238) Welling, M.; Sutton, C. A. Learning in Markov Random Fields with Contrastive Free Energies. In *Artificial Intelligence and Statistics*, 2005.

- (239) Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
- (240) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114* **2013**.
- (241) Zhang, B.; Xiong, D.; Su, J.; Duan, H.; Zhang, M. Variational Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, 521–530.
- (242) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. *arXiv:1511.05644* **2015**.
- (243) LeCun, Y.; Bengio, Y. Convolutional Networks for Images, Speech, and Time Series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, 1995.
- (244) LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional Networks and Applications in Vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, 253–256.
- (245) Lu, H.; Li, Y.; Uemura, T.; Ge, Z.; Xu, X.; He, L.; Serikawa, S.; Kim, H. FDCNet: Filtering Deep Convolutional Network for Marine Organism Classification. *Multimed. Tools Appl.* **2018**, *77*, 21847–21860.
- (246) Srivastava, R. K.; Greff, K.; Schmidhuber, J. Training Very Deep Networks. In *Advances in Neural Information Processing Systems*, 2015, 2377–2385.
- (247) Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; Weinberger, K. Q. Deep Networks with Stochastic Depth. In *European Conference on Computer Vision*, 2016, 646–661.
- (248) Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 4700–4708.
- (249) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
- (250) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2012, 2951–2959.
- (251) Snoek, J.; Rippel, O.; Swersky, K.; Kiros, R.; Satish, N.; Sundaram, N.; Patwary, M.; Prabhat, M.; Adams, R. Scalable Bayesian Optimization Using Deep Neural Networks. In *International Conference on Machine Learning*, 2015, 2171–2180.
- (252) Suganuma, M.; Shirakawa, S.; Nagao, T. A Genetic Programming Approach to Designing Convolutional Neural Network Architectures. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 2017, 497–504.
- (253) Sabour, S.; Frosst, N.; Hinton, G. E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 2017, 3856–3866.
- (254) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014, 2672–2680.
- (255) Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, 2016, 2234–2242.
- (256) Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv:1701.07875* **2017**.
- (257) Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 4681–4690.
- (258) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473* **2014**.
- (259) Graves, A. Generating Sequences with Recurrent Neural Networks. *arXiv:1308.0850* **2013**.
- (260) Luong, M.-T.; Pham, H.; Manning, C. D. Effective Approaches to Attention-Based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, 1412–1421.
- (261) Schneider, G. Mind and Machine in Drug Design. *Nat. Mach. Intell.* **2019**, *1*, 128–130.
- (262) Yu, L.; Wang, S.; Lai, K. K. An Integrated Data Preparation Scheme for Neural Network Data Analysis. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 217–230.
- (263) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons: Weinheim, Germany, 2008.
- (264) Lo, Y.-C.; Rensi, S. E.; Tornig, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1538–1546.
- (265) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Vol. I: Alphabetical Listing/Vol. II: Appendices, References*; John Wiley & Sons: Weinheim, Germany, 2009.
- (266) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. In *Proceedings of the Deep Learning Workshop at NIPS*, 2014, 1–9.
- (267) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*, 2015, 2224–2232.
- (268) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Sci.* **2018**, *4*, 268–276.
- (269) Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.
- (270) Provost, F. Machine Learning from Imbalanced Data Sets 101. In *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*, Austin, TX, July 30–Aug 1, 2000, AAAI, 2000; pp 1–3.
- (271) Brodley, C. E.; Friedl, M. A. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research* **1999**, *11*, 131–167.
- (272) Van Hulse, J.; Khoshgoftaar, T. Knowledge Discovery from Imbalanced and Noisy Data. *Data Knowl. Eng.* **2009**, *68*, 1513–1542.
- (273) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.
- (274) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research* **2002**, *16*, 321–357.
- (275) Maciejewski, T.; Stefanowski, J. Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data. *2011 IEEE Symposium on Computational Intelligence and Data Mining*, Paris, France, Nov 16, 2010; CIDM, 2011; pp 104–111.
- (276) Xiao, X.; Min, J.-L.; Lin, W.-Z.; Liu, Z.; Cheng, X.; Chou, K.-C. iDrug-Target: Predicting the Interactions Between Drug Compounds and Target Proteins in Cellular Networking via Benchmark Dataset Optimization Approach. *J. Biomol. Struct. Dyn.* **2015**, *33*, 2221–2233.
- (277) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283–293.
- (278) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J. Chem. Inf. Model.* **2014**, *54*, 705–712.
- (279) Guha, R.; Schürer, S. C. Utilizing High Throughput Screening Data for Predictive Toxicology Models: Protocols and Application to MLSCN Assays. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 367–384.
- (280) Weis, D. C.; Visco, D. P., Jr; Faulon, J.-L. Data Mining PubChem Using a Support Vector Machine with the Signature Molecular Descriptor: Classification of Factor XIa Inhibitors. *J. Mol. Graphics Modell.* **2008**, *27*, 466–475.
- (281) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

- (282) Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; Rabinovich, A. Training Deep Neural Networks on Noisy Labels with Bootstrapping. *arXiv:1412.6596* 2014.
- (283) Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; Wang, X. Learning from Massive Noisy Labeled Data for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, 2691–2699.
- (284) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discovery* 2010, 9, 273–276.
- (285) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* 2012, 52, 867–881.
- (286) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* 1982, 161, 269–288.
- (287) Lyu, J.; Wang, S.; Baliaus, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaas, E.; Tolmachova, K.; et al. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* 2019, 566, 224–229.
- (288) Sottriffer, C. A.; Gohlke, H.; Klebe, G. Docking into Knowledge-Based Potential Fields: A Comparative Evaluation of DrugScore. *J. Med. Chem.* 2002, 45, 1967–1970.
- (289) Norel, R.; Fischer, D.; Wolfson, H. J.; Nussinov, R. Molecular Surface Recognition by a Computer Vision-Based Technique. *Protein Eng., Des. Sel.* 1994, 7, 39–46.
- (290) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* 1996, 261, 470–489.
- (291) Liu, M.; Wang, S. MCDOCK: A Monte Carlo Simulation Approach to the Molecular Docking Problem. *J. Comput.-Aided Mol. Des.* 1999, 13, 435–451.
- (292) Brooks, B. R.; Brucolieri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* 1983, 4, 187–217.
- (293) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Acc. Chem. Res.* 2017, 50, 302–309.
- (294) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the Development of Universal, Fast and Highly Accurate Docking/Scoring Methods: A Long Way to Go. *Br. J. Pharmacol.* 2008, 153, S7–S26.
- (295) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* 2007, 36, 21–42.
- (296) Ballester, P. J.; Mitchell, J. B. A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* 2010, 26, 1169–1175.
- (297) Brylinski, M. Nonlinear Scoring Functions for Similarity-Based Ligand Docking and Binding Affinity Prediction. *J. Chem. Inf. Model.* 2013, 53, 3097–3112.
- (298) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-Learning Scoring Functions to Improve Structure-Based Binding Affinity Prediction and Virtual Screening. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* 2015, 5, 405–424.
- (299) Kinnings, S. L.; Liu, N.; Tonge, P. J.; Jackson, R. M.; Xie, L.; Bourne, P. E. A Machine Learning-Based Method to Improve Docking Scoring Functions and its Application to Drug Repurposing. *J. Chem. Inf. Model.* 2011, 51, 408–419.
- (300) Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y. ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.* 2013, 53, 592–600.
- (301) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* 2011, 51, 2897–2903.
- (302) Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: A Problem-Centric Review. *AAPS J.* 2012, 14, 133–141.
- (303) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring Functions and Their Evaluation Methods for Protein–Ligand Docking: Recent Advances and Future Directions. *Phys. Chem. Chem. Phys.* 2010, 12, 12899–12908.
- (304) Ma, D.-L.; Chan, D. S.-H.; Leung, C.-H. Drug Repositioning by Structure-Based Virtual Screening. *Chem. Soc. Rev.* 2013, 42, 2130–2141.
- (305) Ballester, P. J.; Schreyer, A.; Blundell, T. L. Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity? *J. Chem. Inf. Model.* 2014, 54, 944–955.
- (306) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Substituting Random Forest for Multiple Linear Regression Improves Binding Affinity Prediction of Scoring Functions: Cyscore as a Case Study. *BMC Bioinformatics* 2014, 15, 291.
- (307) Kramer, C.; Gedeck, P. Leave-Cluster-Out Cross-Validation Is Appropriate for Scoring Functions Derived from Diverse Protein Data Sets. *J. Chem. Inf. Model.* 2010, 50, 1961–1969.
- (308) Ballester, P. J.; Mitchell, J. B. Comments on “leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets”: Significance for the validation of scoring functions. *J. Chem. Inf. Model.* 2011, 51, 1739–1741.
- (309) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* 2014, 54, 2807–2815.
- (310) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest. *Molecules* 2015, 20, 10947–10962.
- (311) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Correcting the Impact of Docking Pose Generation Error on Binding Affinity Prediction. *BMC Bioinformatics* 2016, 17, 308.
- (312) Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein–Ligand Binding Affinity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 2015, 12, 335–347.
- (313) Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A New Open-Source Player in the Drug Discovery Field. *J. Cheminformatics.* 2015, 7, 26.
- (314) Li, H.; Leung, K.-S.; Ballester, P. J.; Wong, M.-H. Istar: A Web Platform for Large-Scale Protein–Ligand Docking. *PLoS One* 2014, 9, e85678.
- (315) Li, H.; Leung, K. S.; Wong, M. H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: the Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* 2015, 34, 115–126.
- (316) Zilian, D.; Sottriffer, C. A. SFCscore RF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* 2013, 53, 1923–1933.
- (317) Liu, Q.; Kwoh, C. K.; Li, J. Binding Affinity Prediction for Protein–Ligand Complexes Based on  $\beta$  Contacts and B Factor. *J. Chem. Inf. Model.* 2013, 53, 3076–3085.
- (318) Li, Y.; Yang, J. Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein–Ligand Interactions. *J. Chem. Inf. Model.* 2017, 57, 1007–1012.
- (319) Li, H.; Peng, J.; Leung, Y.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction. *Biomolecules* 2018, 8, 12.
- (320) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *J. Chem. Inf. Model.* 2009, 49, 1079–1093.
- (321) Ballester, P. J. Machine Learning Scoring Functions Based on Random Forest and Support Vector Regression. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, 2012, 14–25.
- (322) Ashtawy, H. M.; Mahapatra, N. R. BgN-Score and BsN-Score: Bagging and Boosting Based Ensemble Neural Networks Scoring Functions for Accurate Binding Affinity Prediction of Protein–Ligand Complexes. *BMC Bioinformatics* 2015, 16, S8.

- (323) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865–1871.
- (324) Ouyang, X.; Handoko, S. D.; Kwoh, C. K. CScore: A Simple yet Effective Scoring Function for Protein–Ligand Binding Affinity Prediction Using Modified Cmac Learning Architecture. *J. Bioinform. Comput. Biol.* **2011**, *09*, 1–14.
- (325) Cang, Z.; Wei, G. TopologyNet: Topology Based Deep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLoS Comput. Biol.* **2017**, *13*, e1005690.
- (326) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.
- (327) van Dijk, M.; Ter Laak, A. M.; Wichard, J. r. D.; Capoferri, L.; Vermeulen, N. P.; Geerke, D. P. Comprehensive and Automated Linear Interaction Energy Based Binding-Affinity Prediction for Multifarious Cytochrome P450 Aromatase Inhibitors. *J. Chem. Inf. Model.* **2017**, *57*, 2294–2308.
- (328) Cang, Z.; Wei, G. W. Integration of Element Specific Persistent Homology and Machine Learning for Protein-Ligand Binding Affinity Prediction. *Int. J. Numer. Meth. Biomed.* **2018**, *34*, e2914.
- (329) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38*, 169–177.
- (330) Li, L.; Khanna, M.; Jo, I.; Wang, F.; Ashpole, N. M.; Hudmon, A.; Meroueh, S. O. Target-Specific Support Vector Machine Scoring in Structure-Based Virtual Screening: Computational Validation, In Vitro Testing in Kinases, and Effects on Lung Cancer Cell Proliferation. *J. Chem. Inf. Model.* **2011**, *51*, 755–759.
- (331) Xu, D.; Li, L.; Zhou, D.; Liu, D.; Hudmon, A.; Meroueh, S. O. Structure-Based Target-Specific Screening Leads to Small-Molecule CaMKII Inhibitors. *ChemMedChem* **2017**, *12*, 660–677.
- (332) Xu, D.; Meroueh, S. O. Effect of Binding Pose and Modeled Structures on SVMGen and GlideScore Enrichment of Chemical Libraries. *J. Chem. Inf. Model.* **2016**, *56*, 1139–1151.
- (333) Hou, T.; Zhang, W.; Case, D. A.; Wang, W. Characterization of Domain–Peptide Interaction Interface: A Case Study on the Amphiphysin-1 SH3 Domain. *J. Mol. Biol.* **2008**, *376*, 1201–1214.
- (334) Ding, B.; Li, N.; Wang, W. Characterizing Binding of Small Molecules. II. Evaluating the Potency of Small Molecules to Combat Resistance Based on Docking Structures. *J. Chem. Inf. Model.* **2013**, *53*, 1213–1222.
- (335) Li, N.; Ainsworth, R. I.; Wu, M.; Ding, B.; Wang, W. MIEC-SVM: Automated Pipeline for Protein Peptide/Ligand Interaction Prediction. *Bioinformatics* **2016**, *32*, 940–942.
- (336) Sun, H.; Pan, P.; Tian, S.; Xu, L.; Kong, X.; Li, Y.; Li, D.; Hou, T. Constructing and Validating High-Performance MIEC-SVM Models in Virtual Screening for Kinases: A Better Way for Actives Discovery. *Sci. Rep.* **2016**, *6*, 24817.
- (337) Leong, M. K.; Syu, R.-G.; Ding, Y.-L.; Weng, C.-F. Prediction of N-Methyl-D-Aspartate Receptor GluN1-Ligand Binding Affinity by a Novel SVM-Pose/SVM-Score Combinatorial Ensemble Docking Scheme. *Sci. Rep.* **2017**, *7*, 40053.
- (338) Yan, Y.; Wang, W.; Sun, Z.; Zhang, J. Z.; Ji, C. Protein–Ligand Empirical Interaction Components for Virtual Screening. *J. Chem. Inf. Model.* **2017**, *57*, 1793–1806.
- (339) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. r. Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 710–716.
- (340) Xie, Q.-Q.; Zhong, L.; Pan, Y.-L.; Wang, X.-Y.; Zhou, J.-P.; Diwu, L.; Huang, Q.; Wang, Y.-L.; Yang, L.-L.; Xie, H.-Z.; et al. Combined SVM-Based and Docking-Based Virtual Screening for Retrieving Novel Inhibitors of c-Met. *Eur. J. Med. Chem.* **2011**, *46*, 3675–3680.
- (341) Meslamani, J.; Bhajun, R.; Martz, F.; Rognan, D. Computational Profiling of Bioactive Compounds Using a Target-Dependent Composite Workflow. *J. Chem. Inf. Model.* **2013**, *53*, 2322–2333.
- (342) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. *Sci. Rep.* **2017**, *7*, 46710.
- (343) Soufan, O.; Ba-Alawi, W.; Afeef, M.; Essack, M.; Kalnis, P.; Bajic, V. B. DRABAL: Novel Method to Mine Large High-Throughput Screening Assays Using Bayesian Active Learning. *J. Cheminform.* **2016**, *8*, 64.
- (344) Arciniega, M.; Lange, O. F. Improvement of Virtual Screening Results by Docking Data Feature Analysis. *J. Chem. Inf. Model.* **2014**, *54*, 1401–1411.
- (345) Sunseri, J.; Ragoza, M.; Collins, J.; Koes, D. R. A D3R Prospective Evaluation of Machine Learning for Protein-Ligand Scoring. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 761–771.
- (346) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2013**, *35*, 1798–1828.
- (347) Pereira, J. C.; Caffarena, E. R. l.; dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* **2016**, *56*, 2495–2506.
- (348) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
- (349) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1044.
- (350) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
- (351) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in Drug Design-A Review. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115.
- (352) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.
- (353) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: a Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
- (354) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077–2088.
- (355) Marchese Robinson, R. L.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792.
- (356) Chen, J. J. F.; Visco, D. P., Jr Developing an in Silico Pipeline for Faster Drug Candidate Discovery: Virtual High Throughput Screening with the Signature Molecular Descriptor Using Support Vector Machine Models. *Chem. Eng. Sci.* **2017**, *159*, 31–42.
- (357) Chen, J. J. F.; Visco, D. P., Jr Identifying Novel Factor XIIa Inhibitors with PCA-GA-SVM Developed vHTS Models. *Eur. J. Med. Chem.* **2017**, *140*, 31–41.
- (358) Fang, X.; Bagui, S.; Bagui, S. Improving Virtual Screening Predictive Accuracy of Human Kallikrein5 Inhibitors Using Machine Learning Models. *Comput. Biol. Chem.* **2017**, *69*, 110–119.
- (359) Alvarsson, J.; Lampa, S.; Schaal, W.; Andersson, C.; Wikberg, J. E.; Spjuth, O. Large-Scale Ligand-Based Predictive Modelling Using Support Vector Machines. *J. Cheminform.* **2016**, *8*, 39.
- (360) Alvarsson, J.; Eklund, M.; Andersson, C.; Carlsson, L.; Spjuth, O.; Wikberg, J. E. Benchmarking Study of Parameter Variation When Using Signature Fingerprints Together with Support Vector Machines. *J. Chem. Inf. Model.* **2014**, *54*, 3211–3217.
- (361) Rensi, S. E.; Altman, R. B. Shallow Representation Learning via Kernel PCA Improves QSAR Modelability. *J. Chem. Inf. Model.* **2017**, *57*, 1859–1867.
- (362) Clark, A. M.; Dole, K.; Coulon-Spektor, A.; McNutt, A.; Grass, G.; Freundlich, J. S.; Reynolds, R. C.; Ekins, S. Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. *J. Chem. Inf. Model.* **2015**, *55*, 1231–1245.

- (363) Ekins, S.; Reynolds, R. C.; Kim, H.; Koo, M.-S.; Ekonomidis, M.; Talaue, M.; Paget, S. D.; Woolhiser, L. K.; Lenaerts, A. J.; Bunin, B. A.; et al. Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery. *Chem. Biol.* **2013**, *20*, 370–378.
- (364) Liu, L.-l.; Lu, J.; Lu, Y.; Zheng, M.-y.; Luo, X.-m.; Zhu, W.-l.; Jiang, H.-l.; Chen, K.-x. Novel Bayesian Classification Models for Predicting Compounds Blocking hERG Potassium Channels. *Acta Pharmacol. Sin.* **2014**, *35*, 1093–1102.
- (365) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463–4470.
- (366) Fang, J.; Li, Y.; Liu, R.; Pang, X.; Li, C.; Yang, R.; He, Y.; Lian, W.; Liu, A.-L.; Du, G.-H. Discovery of Multitarget-Directed Ligands Against Alzheimer's Disease Through Systematic Prediction of Chemical–Protein Interactions. *J. Chem. Inf. Model.* **2015**, *55*, 149–164.
- (367) Fang, J.; Wang, L.; Li, Y.; Lian, W.; Pang, X.; Wang, H.; Yuan, D.; Wang, Q.; Liu, A.-L.; Du, G.-H. AlzhCPI: A Knowledge Base for Predicting Chemical–Protein Interactions Towards Alzheimer's Disease. *PLoS One* **2017**, *12*, e0178347.
- (368) Renault, N.; Laurent, X.; Farce, A.; El Bakali, J.; Mansouri, R.; Gervois, P.; Millet, R.; Desreumaux, P.; Furman, C.; Chavatte, P. Virtual Screening of CB2 Receptor Agonists from Bayesian Network and High-Throughput Docking: Structural Insights into Agonist-Modulated GPCR Features. *Chem. Biol. Drug Des.* **2013**, *81*, 442–454.
- (369) Bender, A.; Mussa, H. Y.; Glen, R. C. Screening for Dihydrofolate Reductase Inhibitors Using MOLPRINT 2D, a Fast Fragment-Based Method Employing the Naive Bayesian Classifier: Limitations of the Descriptor and the Importance of Balanced Chemistry in Training and Test Sets. *J. Biomol. Screening* **2005**, *10*, 658–666.
- (370) Vijayan, R.; Bera, I.; Prabu, M.; Saha, S.; Ghoshal, N. Combinatorial Library Enumeration and Lead Hopping Using Comparative Interaction Fingerprint Analysis and Classical 2D QSAR Methods for Seeking Novel GABA $\alpha$  Modulators. *J. Chem. Inf. Model.* **2009**, *49*, 2498–2511.
- (371) Prathipati, P.; Ma, N. L.; Keller, T. H. Global Bayesian Models for the Prioritization of Antitubercular Agents. *J. Chem. Inf. Model.* **2008**, *48*, 2362–2370.
- (372) Singh, N.; Chaudhury, S.; Liu, R.; AbdulHameed, M. D. M.; Tawa, G.; Wallqvist, A. QSAR Classification Model for Antibacterial Compounds and Its Use in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52*, 2559–2569.
- (373) Hu, Y.; Unwalla, R.; Denny, R. A.; Bikker, J.; Di, L.; Humblet, C. Development of QSAR Models for Microsomal Stability: Identification of Good and Bad Structural Features for Rat, Human and Mouse Microsomal Stability. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 23–35.
- (374) Pan, Y.; Li, L.; Kim, G.; Ekins, S.; Wang, H.; Swaan, P. W. Identification and Validation of Novel Human Pregnane X Receptor Activators Among Prescribed Drugs via Ligand-Based Virtual Screening. *Drug Metab. Dispos.* **2011**, *39*, 337–344.
- (375) Shi, H.; Tian, S.; Li, Y.; Li, D.; Yu, H.; Zhen, X.; Hou, T. Absorption, Distribution, Metabolism, Excretion, and Toxicity Evaluation in Drug Discovery. 14. Prediction of Human Pregnane X Receptor Activators by Using Naive Bayesian Classification Technique. *Chem. Res. Toxicol.* **2015**, *28*, 116–125.
- (376) AbdulHameed, M. D. M.; Ippolito, D. L.; Wallqvist, A. Predicting Rat and Human Pregnane X Receptor Activators Using Bayesian Classification Models. *Chem. Res. Toxicol.* **2016**, *29*, 1729–1740.
- (377) Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-Villanueva, W.; Castro-Bleda, M. J.; Villanueva-Pareja, A. Artificial Neural Networks and Linear Discriminant Analysis: A Valuable Combination in the Selection of New Antibacterial Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1031–1041.
- (378) Sabet, R.; Fassihi, A.; Hemmateenejad, B.; Saghaei, L.; Miri, R.; Gholami, M. Computer-Aided Design of Novel Antibacterial 3-Hydroxypyridine-4-Ones: Application of QSAR Methods Based on the MOLMAP Approach. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 349–361.
- (379) Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Pante, N.; Hancock, R. E.; Cherkasov, A. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* **2009**, *52*, 2006–2015.
- (380) Cherkasov, A.; Hilpert, K.; Jenssen, H.; Fjell, C. D.; Waldbrook, M.; Mullaly, S. C.; Volkmer, R.; Hancock, R. E. Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective Against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chem. Biol.* **2009**, *4*, 65–74.
- (381) Torrent, M.; Andreu, D.; Nogués, V. M.; Boix, E. Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS One* **2011**, *6*, e16968.
- (382) Douali, L.; Villemain, D.; Cherqaoui, D. Neural Networks: Accurate Nonlinear QSAR Model for HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1200–1207.
- (383) Myint, K.-Z.; Wang, L.; Tong, Q.; Xie, X.-Q. Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. *Mol. Pharmaceutics* **2012**, *9*, 2912–2923.
- (384) Sardari, S.; Kohanzad, H.; Ghavami, G. Artificial Neural Network Modeling of Antimycobacterial Chemical Space to Introduce Efficient Descriptors Employed for Drug Design. *Chemom. Intell. Lab. Syst.* **2014**, *130*, 151–158.
- (385) Hu, L.; Chen, G.; Chau, R. M.-W. A Neural Networks-Based Drug Discovery Approach and Its Application for Designing Aldose Reductase Inhibitors. *J. Mol. Graphics Modell.* **2006**, *24*, 244–253.
- (386) Patra, J. C.; Chua, B. H. Artificial Neural Network-Based Drug Design for Diabetes Mellitus Using Flavonoids. *J. Comput. Chem.* **2011**, *32*, 555–567.
- (387) Khatri, N.; Lather, V.; Madan, A. Diverse Classification Models for Anti-Hepatitis C Virus Activity of Thiourea Derivatives. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 13–21.
- (388) Jaén-Oltra, J.; Salabert-Salvador, M. T.; García-March, F. J.; Pérez-Giménez, F.; Tomás-Vert, F. Artificial Neural Network Applied to Prediction of Fluorquinolone Antibacterial Activity by Topological Methods. *J. Med. Chem.* **2000**, *43*, 1143–1148.
- (389) Geanes, A. R.; Cho, H. P.; Nance, K. D.; McGowan, K. M.; Conn, P. J.; Jones, C. K.; Meiler, J.; Lindsley, C. W. Ligand-Based Virtual Screen for the Discovery of Novel M5 Inhibitor Chemotypes. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 4487–4491.
- (390) Vina, D.; Uriarte, E.; Orallo, F.; González-Díaz, H. Alignment-Free Prediction of a Drug–Target Complex Network Based on Parameters of Drug Connectivity and Protein Sequence of Receptors. *Mol. Pharmaceutics* **2009**, *6*, 825–835.
- (391) Prado-Prado, F. J.; Ubeira, F. M.; Borges, F.; González-Díaz, H. Unified QSAR & Network-Based Computational Chemistry Approach to Antimicrobials. II. Multiple Distance and Triadic Census Analysis of Antiparasitic Drugs Complex Networks. *J. Comput. Chem.* **2010**, *31*, 164–173.
- (392) Speck-Planche, A.; V Kleandrova, V.; Luan, F.; Cordeiro, N. D. Chemoinformatics in Multi-Target Drug Discovery for Anti-Cancer Therapy: In Silico Design of Ptent and Vrsatile Ati-Bain Tmor Aents. *Anticancer Agents Med. Chem.* **2012**, *12*, 678–685.
- (393) Speck-Planche, A.; Kleandrova, V. V.; Cordeiro, M. N. D. New Insights Toward the Discovery of Antibacterial Agents: Multi-Tasking QSBER Model for the Simultaneous Prediction of Anti-Tuberculosis Activity and Toxicological Profiles of Drugs. *Eur. J. Pharm. Sci.* **2013**, *48*, 812–818.
- (394) Tenorio-Borroto, E.; Rivas, C. G. P.; Chagoyán, J. C. V.; Castañedo, N.; Prado-Prado, F. J.; García-Mera, X.; González-Díaz, H. ANN Multiplexing Model of Drugs Effect on Macrophages; Theoretical and Flow Cytometry Study on the Cytotoxicity of the Anti-Microbial Drug G1 in Spleen. *Bioorg. Med. Chem.* **2012**, *20*, 6181–6194.

- (395) Tenorio-Borroto, E.; G Penuelas-Rivas, C.; C Vasquez-Chagoyan, J.; J Prado-Prado, F.; García-Mera, X.; Gonzalez-Diaz, H. Immunotoxicity, Flow Cytometry, and Chemoinformatics: Review, Bibliometric Analysis, and New QSAR Model of Drug Effects over Macrophages. *Curr. Top. Med. Chem.* **2012**, *12*, 1815–1833.
- (396) Tenorio-Borroto, E.; García-Mera, X.; G Penuelas-Rivas, C.; C Vasquez-Chagoyan, J.; J Prado-Prado, F.; Castanedo, N.; Gonzalez-Diaz, H. Entropy Model for Multiplex Drug-Target Interaction Endpoints of Drug Immunotoxicity. *Curr. Top. Med. Chem.* **2013**, *13*, 1636–1649.
- (397) Tenorio-Borroto, E.; Peñuelas-Rivas, C. G.; Vásquez-Chagoyán, J. C.; Castañedo, N.; Prado-Prado, F. J.; García-Mera, X.; González-Díaz, H. Model for High-Throughput Screening of Drug Immunotoxicity—Study of the Anti-Microbial G1 over Peritoneal Macrophages Using Flow Cytometry. *Eur. J. Med. Chem.* **2014**, *72*, 206–220.
- (398) Speck-Planche, A.; Cordeiro, M. Simultaneous Modeling of Antimycobacterial Activities and ADMET Profiles: A Chemoinformatic Approach to Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 1656–1665.
- (399) Speck-Planche, A.; Cordeiro, M. Simultaneous Virtual Prediction of Anti-*Escherichia coli* Activities and ADMET Profiles: A Chemoinformatic Complementary Approach for High-Throughput Screening. *ACS Comb. Sci.* **2014**, *16*, 78–84.
- (400) Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. I. Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb. Sci.* **2017**, *19*, 501–512.
- (401) Speck-Planche, A.; Cordeiro, M. N. D. Fragment-Based in Silico Modeling of Multi-Target Inhibitors Against Breast Cancer-Related Proteins. *Mol. Diversity* **2017**, *21*, 511–523.
- (402) Kleandrova, V. V.; Russo, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N. I. Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb. Sci.* **2016**, *18*, 490–498.
- (403) Lenselink, E. B.; Dijke, N.; Bongers, B.; Papadatos, G.; Vlijmen, H. W.; Kowalczyk, W.; IJzerman, A. P.; Westen, G. J. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform.* **2017**, *9*, 45.
- (404) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.
- (405) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.
- (406) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv:1502.02072* **2015**.
- (407) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490–2504.
- (408) Winkler, D. A.; Le, T. C. Corrigendum: Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Mol. Inf.* **2017**, *36*, 1781141.
- (409) Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J. Deep-Learning: Investigating Deep Neural Networks Hyper-Parameters and Comparison of Performance to Shallow Methods for Modeling Bioactivity Data. *J. Cheminformatics.* **2017**, *9*, 42.
- (410) Mendenhall, J.; Meiler, J. Improving Quantitative Structure–Activity Relationship Models Using Artificial Neural Networks Trained with Dropout. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 177–189.
- (411) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175.
- (412) Korkmaz, S.; Zararsiz, G.; Gokseluk, D. Drug/Nondrug Classification Using Support Vector Machines with Various Feature Selection Strategies. *Comput. Methods Programs Biomed.* **2014**, *117*, 51–60.
- (413) García-Sosa, A. T.; Oja, M.; Hetényi, C.; Maran, U. DrugLogit: Logistic Discrimination Between Drugs and Nondrugs Including Disease-Specificity by Assigning Probabilities Based on Molecular Properties. *J. Chem. Inf. Model.* **2012**, *52*, 2165–2180.
- (414) Smusz, S.; Kurczab, R.; Satala, G.; Bojarski, A. J. Fingerprint-Based Consensus Virtual Screening Towards Structurally New 5-HT6R ligands. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 1827–1830.
- (415) Balfer, J.; Bajorath, J. r. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55*, 1136–1147.
- (416) Liu, J.; Ning, X. Multi-Assay-Based Compound Prioritization via Assistance Utilization: A Machine Learning Framework. *J. Chem. Inf. Model.* **2017**, *57*, 484–498.
- (417) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naive Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1945–1956.
- (418) Rogers, D.; Brown, R. D.; Hahn, M. Using Extended-Connectivity Fingerprints with Laplacian-Modified Bayesian Analysis in High-Throughput Screening Follow-Up. *J. Biomol. Screening* **2005**, *10*, 682–686.
- (419) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of Random Forest and Pipeline Pilot Naïve Bayes in Prospective QSAR Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792–803.
- (420) Zhang, Y.; Jiao, Y.; Xiong, X.; Liu, H.; Ran, T.; Xu, J.; Lu, S.; Xu, A.; Pan, J.; Qiao, X.; et al. Fragment Virtual Screening Based on Bayesian Categorization for Discovering Novel VEGFR-2 Scaffolds. *Mol. Divers.* **2015**, *19*, 895–913.
- (421) Ding, Q.; Li, C.; Wang, L.; Li, Y.; Zhou, H.; Gu, Q.; Xu, J. Identifying Farnesoid X Receptor Agonists by Naïve Bayesian and Recursive Partitioning Approaches. *MedChemComm* **2015**, *6*, 1393–1403.
- (422) Li, Y.; Wang, L.; Liu, Z.; Li, C.; Xu, J.; Gu, Q.; Xu, J. Predicting Selective Liver X Receptor  $\beta$  Agonists Using Multiple Machine Learning Methods. *Mol. BioSyst.* **2015**, *11*, 1241–1250.
- (423) Cai, C.; Wu, Q.; Luo, Y.; Ma, H.; Shen, J.; Zhang, Y.; Yang, L.; Chen, Y.; Wen, Z.; Wang, Q. In Silico Prediction of ROCK II Inhibitors by Different Classification Approaches. *Mol. Diversity* **2017**, *21*, 791–807.
- (424) Wang, L.; Li, Y.; Xu, M.; Pang, X.; Liu, Z.; Tan, W.; Xu, J. Chemical Fragment-Based CDK4/6 Inhibitors Prediction and Web Server. *RSC Adv.* **2016**, *6*, 16972–16981.
- (425) Li, L.; Le, X.; Wang, L.; Gu, Q.; Zhou, H.; Xu, J. Discovering New DNA Gyrase Inhibitors Using Machine Learning Approaches. *RSC Adv.* **2015**, *5*, 105600–105608.
- (426) Pan, Y.; Chothe, P. P.; Swaan, P. W. Identification of Novel Breast Cancer Resistance Protein (BCRP) Inhibitors by Virtual Screening. *Mol. Pharmaceutics* **2013**, *10*, 1236–1248.
- (427) Montanari, F.; Cseke, A.; Wlcek, K.; Ecker, G. F. Virtual Screening of DrugBank Reveals Two Drugs as New BCRP Inhibitors. *J. Biomol. Screen.* **2017**, *22*, 86–93.
- (428) Cai, J.; Li, C.; Liu, Z.; Du, J.; Ye, J.; Gu, Q.; Xu, J. Predicting DPP-IV Inhibitors with Machine Learning Approaches. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 393–402.
- (429) Fang, J.; Yang, R.; Gao, L.; Zhou, D.; Yang, S.; Liu, A.-I.; Du, G.-h. Predictions of BuChE Inhibitors Using Support Vector Machine and Naïve Bayesian Classification Techniques in Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53*, 3009–3020.
- (430) Wicht, K. J.; Combrinck, J. M.; Smith, P. J.; Egan, T. J. Bayesian Models Trained with HTS Data for Predicting  $\beta$ -Haematin Inhibition and in Vitro Antimalarial Activity. *Bioorg. Med. Chem.* **2015**, *23*, 5210–5217.
- (431) Ekins, S.; Perryman, A. L.; Clark, A. M.; Reynolds, R. C.; Freundlich, J. S. Machine Learning Model Analysis and Data Visualization with Small Molecules Tested in a Mouse Model of

- Mycobacterium Tuberculosis Infection (2014–2015). *J. Chem. Inf. Model.* **2016**, *56*, 1332–1343.
- (432) Ekins, S.; Godbole, A. A.; Kéri, G.; Orfi, L.; Pato, J.; Bhat, R. S.; Verma, R.; Bradley, E. K.; Nagaraja, V. Machine Learning and Docking Models for Mycobacterium Tuberculosis Topoisomerase I. *Tuberculosis* **2017**, *103*, 52–60.
- (433) Ekins, S.; Freundlich, J. S.; Reynolds, R. C. Are Bigger Data Sets Better for Machine Learning? Fusing Single-Point and Dual-Event Dose Response Data for Mycobacterium Tuberculosis. *J. Chem. Inf. Model.* **2014**, *54*, 2157–2165.
- (434) Clark, A. M.; Ekins, S. Open Source Bayesian Models. 2. Mining a “Big Dataset” to Create and Validate Models with ChEMBL. *J. Chem. Inf. Model.* **2015**, *55*, 1246–1260.
- (435) Clark, A. M.; Dole, K.; Ekins, S. Open Source Bayesian Models. 3. Composite Models for Prediction of Binned Responses. *J. Chem. Inf. Model.* **2016**, *56*, 275–285.
- (436) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942–4951.
- (437) Abdo, A.; Salim, N. Similarity-Based Virtual Screening with a Bayesian Inference Network. *ChemMedChem* **2009**, *4*, 210–218.
- (438) Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012–1020.
- (439) Abdo, A.; Leclère, V.; Jacques, P.; Salim, N.; Pupin, M. Prediction of New Bioactive Molecules Using a Bayesian Belief Network. *J. Chem. Inf. Model.* **2014**, *54*, 30–36.
- (440) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473.
- (441) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer: Berlin, Germany, 2009.
- (442) Schneider, P.; Schneider, G. Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR Comb. Sci.* **2003**, *22*, 713–718.
- (443) Reker, D.; Seet, M.; Pillong, M.; Koch, C. P.; Schneider, P.; Witschel, M. C.; Rottmann, M.; Freymond, C.; Brun, R.; Schweizer, B.; et al. Deorphaning Pyrrolopyrazines as Potent Multi-Target Antimalarial Agents. *Angew. Chem.-Int. Ed.* **2014**, *53*, 7079–7084.
- (444) Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Mönch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; et al. Revealing the Macromolecular Targets of Complex Natural Products. *Nat. Chem.* **2014**, *6*, 1072–1078.
- (445) Schneider, G.; Reker, D.; Chen, T.; Hauenstein, K.; Schneider, P.; Altmann, K. H. Deorphaning the Macromolecular Targets of the Natural Anticancer Compound Dolichulide. *Angew. Chem., Int. Ed.* **2016**, *55*, 12408–12411.
- (446) Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8*, 531–541.
- (447) Rodrigues, T.; Reker, D.; Kunze, J.; Schneider, P.; Schneider, G. Revealing the Macromolecular Targets of Fragment-Like Natural Products. *Angew. Chem., Int. Ed.* **2015**, *54*, 10516–10520.
- (448) Hecht-Nielsen, R. Counterpropagation Networks. *Appl. Opt.* **1987**, *26*, 4979–4984.
- (449) Merk, D.; Grisoni, F.; Schaller, K.; Friedrich, L.; Schneider, G. Discovery of Novel Molecular Frameworks of Farnesoid X Receptor Modulators by Ensemble Machine Learning. *ChemistryOpen* **2019**, *8*, 7–14.
- (450) Conant, G.; Ahadian, S.; Zhao, Y.; Radisic, M. Kinase Inhibitor Screening Using Artificial Neural Networks and Engineered Cardiac Biowires. *Sci. Rep.* **2017**, *7*, 11807.
- (451) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.
- (452) Halls, M. D.; Giesen, D. J.; Hughes, T. F.; Goldberg, A.; Cao, Y. High-Throughput Quantum Chemistry and Virtual Screening for OLED Material Components. In *Organic Light Emitting Materials and Devices XVII*, 2013, p 882926.
- (453) Halls, M. D.; Tasaki, K. High-Throughput Quantum Chemistry and Virtual Screening for Lithium Ion Battery Electrolyte Additives. *J. Power Sources* **2010**, *195*, 1472–1478.
- (454) Segall, M. Advances in Multiparameter Optimization Methods for De Novo Drug Design. *Expert Opin. Drug Discovery* **2014**, *9*, 803–817.
- (455) Schneider, G.; Fechner, U. Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.
- (456) Schneider, G. Future De Novo Drug Design. *Mol. Inf.* **2014**, *33*, 397–402.
- (457) Böhm, H.-J. The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
- (458) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Developments in the De Novo Design of Molecules. *J. Chem. Inf. Model.* **1994**, *34*, 207–217.
- (459) Ruddigkeit, L.; Blum, L. C.; Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2013**, *53*, 56–65.
- (460) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven De Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8*, e1002380.
- (461) Schneider, G.; Geppert, T.; Hartenfeller, M.; Reisen, F.; Klenner, A.; Reutlinger, M.; Hähnke, V.; Hiss, J. A.; Zettl, H.; Keppner, S. Reaction-Driven De Novo Design, Synthesis and Testing of Potential Type II Kinase Inhibitors. *Future Medicinal Chemistry* **2011**, *3*, 415–424.
- (462) Friedrich, L.; Rodrigues, T.; Neuhaus, C. S.; Schneider, P.; Schneider, G. From Complex Natural Products to Simple Synthetic Mimetics by Computational De Novo Design. *Angew. Chem., Int. Ed.* **2016**, *55*, 6789–6792.
- (463) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguez, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; et al. Automated Design of Ligands to Polypharmacological Profiles. *Nature* **2012**, *492*, 215–220.
- (464) Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian Molecular Design with a Chemical Language Model. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 379–391.
- (465) Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). *J. Chem. Inf. Model.* **2016**, *56*, 286–299.
- (466) Churchwell, C. J.; Rintoul, M. D.; Martin, S.; Visco, D. P., Jr; Kotu, A.; Larson, R. S.; Sillerud, L. O.; Brown, D. C.; Faulon, J.-L. The Signature Molecular Descriptor: 3. Inverse-Quantitative Structure–Activity Relationship of ICAM-1 Inhibitory Peptides. *J. Mol. Graphics Modell.* **2004**, *22*, 263–273.
- (467) Wong, W. W.; Burkowski, F. J. A Constructive Approach for Discovering New Drug Leads: Using a Kernel Methodology for the Inverse-QSAR Problem. *J. Cheminform.* **2009**, *1*, 1–27.
- (468) Miyao, T.; Arakawa, M.; Funatsu, K. Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Mol. Inf.* **2010**, *29*, 111–125.
- (469) Takeda, S.; Kaneko, H.; Funatsu, K. Chemical-Space-Based De Novo Design Method To Generate Drug-Like Molecules. *J. Chem. Inf. Model.* **2016**, *56*, 1885–1893.
- (470) Mishima, K.; Kaneko, H.; Funatsu, K. Development of a New De Novo Design Algorithm for Exploring Chemical Space. *Mol. Inf.* **2014**, *33*, 779–789.
- (471) Kawai, K.; Nagata, N.; Takahashi, Y. De Novo Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach. *J. Chem. Inf. Model.* **2014**, *54*, 49–56.
- (472) Dey, F.; Caflisch, A. Fragment-Based De Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.
- (473) Fechner, U.; Schneider, G. Flux(1): A Virtual Synthesis Scheme for Fragment-Based De Novo Design. *J. Chem. Inf. Model.* **2006**, *46*, 699–707.

- (474) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8*, 10883–10890.
- (475) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. DruGAN: An Advanced Generative Adversarial Autoencoder Model for De Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.
- (476) Jaques, N.; Gu, S.; Bahdanau, D.; Hernández-Lobato, J. M.; Turner, R. E.; Eck, D. Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-Control. In *Proceedings of the 34th International Conference on Machine Learning*, 2017, 1645–1654.
- (477) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design Through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9*, 48.
- (478) Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.
- (479) Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017, 2852–2858.
- (480) Harel, S.; Radinsky, K. Prototype-Based Compound Discovery Using Deep Generative Models. *Mol. Pharmaceutics* **2018**, *15*, 4406–4416.
- (481) Yuan, W.; Jiang, D.; Nambiar, D. K.; Liew, L. P.; Hay, M. P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.-T.; Tibshirani, R.; et al. Chemical Space Mimicry for Drug Discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875–882.
- (482) Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning Artificial Intelligence on the De Novo Design of Natural-Product-Inspired Retinoid X Receptor Modulators. *Commun. Chem.* **2018**, *1*, 68.
- (483) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
- (484) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 Chemical Space Using Deep Generative Models. *J. Cheminform.* **2019**, *11*, 20.
- (485) Pogany, P.; Arad, N.; Genway, S.; Pickett, S. D. De Novo Molecule Design by Translating from Reduced Graphs to SMILES. *J. Chem. Inf. Model.* **2019**, *59*, 1136–1146.
- (486) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for De Novo Molecular Design. *J. Cheminform.* **2018**, *10*, 31.
- (487) Preuer, K.; Renz, P.; Unterthiner, T.; Hochreiter, S.; Klambauer, G. n. Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58*, 1736–1741.
- (488) Fréchet, M. Sur La Distance De Deux Lois De Probabilité. *C. R. Acad. Sci. Paris.* **1957**, *244*, 689–692.
- (489) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9*, 5441–5451.
- (490) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971**, *71*, 525–616.
- (491) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98.
- (492) Dayan, P.; Balleine, B. W. Reward, Motivation, and Reinforcement Learning. *Neuron* **2002**, *36*, 285–298.
- (493) Kaelbling, L. P.; Littman, M. L.; Moore, A. W. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* **1996**, *4*, 237–285.
- (494) Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; Meger, D. Deep Reinforcement Learning that Matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- (495) Liu, X.; Ye, K.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. An Exploration Strategy Improves the Diversity of De Novo Ligands Using Deep Reinforcement Learning—A Case for the Adenosine A2A Receptor. *J. Cheminform.* **2019**, *11*, 25.
- (496) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De Novo Drug Design. *Sci. Adv.* **2018**, *4*, eaap7885.
- (497) Joulin, A.; Mikolov, T. Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. In *Advances in Neural Information Processing Systems*, 2015, 190–198.
- (498) Awale, M.; Sirockin, F.; Stiefl, N.; Reymond, J.-L. Drug Analogs from Fragment Based Long Short-Term Memory Generative Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1347–1356.
- (499) Visini, R.; Awale, M.; Reymond, J.-L. Fragment Database FDB-17. *J. Chem. Inf. Model.* **2017**, *57*, 700–709.
- (500) Gupta, A.; Müller, A. T.; Huisman, B. J.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.
- (501) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdagyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for De Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398–4405.
- (502) Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquier, D.; Wichard, J. De Novo Generation of Hit-Like Molecules from Gene Expression Signatures Using Artificial Intelligence. *chemRxiv:7294388* **2018**, DOI: [10.26434/chemrxiv.7294388.v1](https://doi.org/10.26434/chemrxiv.7294388.v1).
- (503) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171*, 1437–1452.
- (504) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing Distributions over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-Design Chemistry (ORGANIC). *chemRxiv:5309668* **2017**, DOI: [10.26434/chemrxiv.5309668.v3](https://doi.org/10.26434/chemrxiv.5309668.v3).
- (505) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sánchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for De Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.
- (506) Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Colmenarejo, S. G.; Grefenstette, E.; Ramalho, T.; Agapiou, J. Hybrid Computing Using a Neural Network with Dynamic External Memory. *Nature* **2016**, *538*, 471–476.
- (507) Muegge, I. Selection Criteria for Drug-Like Compounds. *Med. Res. Rev.* **2003**, *23*, 302–321.
- (508) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular De Novo Design. *Mol. Pharmaceutics* **2018**, *15*, 4386–4397.
- (509) Yoshikawa, N.; Terayama, K.; Honma, T.; Oono, K.; Tsuda, K. Population-Based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **2018**, *47*, 1431–1434.
- (510) Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, 2017, 1945–1954.
- (511) Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *chemRxiv:7097960*; **2018**, DOI: [10.26434/chemrxiv.7097960.v1](https://doi.org/10.26434/chemrxiv.7097960.v1).
- (512) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113.
- (513) De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *arXiv:1805.11973* **2018**.
- (514) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Warchoł, M. Mol-CycleGAN-a Generative Model for Molecular Optimization. *arXiv:1902.02119* **2019**.
- (515) Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In

- Proceedings of the IEEE International Conference on Computer Vision, 2017; pp 2223–2232.
- (516) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *International Conference on Machine Learning*, 2018; pp 2328–2337.
- (517) Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In *International Conference on Artificial Neural Networks*, 2018; pp 412–422.
- (518) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective De Novo Drug Design with Conditional Graph Generative Model. *J. Cheminform.* **2018**, *10*, 33.
- (519) Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained Graph Variational Autoencoders for Molecule Design. In *Advances in Neural Information Processing Systems*, 2018; pp 7806–7815.
- (520) You, J.; Liu, B.; Ying, R.; Pande, V.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. In *Advances in Neural Information Processing Systems*, 2018; pp 6410–6421.
- (521) You, J.; Ying, R.; Ren, X.; Hamilton, W.; Leskovec, J. GraphRNN: Generating Realistic Graphs with Deep Auto-Regressive Models. In *International Conference on Machine Learning*, 2018; pp 5694–5703.
- (522) Xu, Z.; Wang, S.; Zhu, F.; Huang, J. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2017; pp 285–294.
- (523) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: An Efficient Python Library for De Novo Molecular Generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.
- (524) Schneider, G.; Clark, D. E. Automated De Novo Drug Design—“Are We Nearly There Yet?”. *Angew. Chem., Int. Ed.* **2019**, *58*, DOI: 10.1002/anie.201814681.
- (525) Schneider, P.; Schneider, G. De Novo Design at the Edge of Chaos: Miniperspective. *J. Med. Chem.* **2016**, *59*, 4077–4086.
- (526) Nobeli, I.; Favia, A. D.; Thornton, J. M. Protein Promiscuity and Its Implications for Biotechnology. *Nat. Biotechnol.* **2009**, *27*, 157–167.
- (527) Hopkins, A. L. Drug Discovery: Predicting Promiscuity. *Nature* **2009**, *462*, 167–168.
- (528) Morgan, P.; Van Der Graaf, P. H.; Arrowsmith, J.; Feltner, D. E.; Drummond, K. S.; Wegner, C. D.; Street, S. D. Can the Flow of Medicines be Improved? Fundamental Pharmacokinetic and Pharmacological Principles Toward Improving Phase II Survival. *Drug Discovery Today* **2012**, *17*, 419–424.
- (529) Cumming, J. G.; Davis, A. M.; Muresan, S.; Haeberlein, M.; Chen, H. Chemical Predictive Modelling to Improve Compound Quality. *Nat. Rev. Drug Discovery* **2013**, *12*, 948–963.
- (530) Chung, T. D.; Terry, D. B.; Smith, L. H. *In Vitro and In Vivo Assessment of ADME and PK Properties During Lead Selection and Lead Optimization—Guidelines, Benchmarks and Rules of Thum*; Eli Lilly & Company and the National Center for Advancing Translational Sciences: Bethesda, MD, USA, 2015.
- (531) Wang, Y.; Xing, J.; Xu, Y.; Zhou, N.; Peng, J.; Xiong, Z.; Liu, X.; Luo, X.; Luo, C.; Chen, K. In Silico ADME/T Modelling for Rational Drug Design. *Q. Rev. Biophys.* **2015**, *48*, 488–515.
- (532) Van De Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.
- (533) Sarkar, A.; Kellogg, G. E. Hydrophobicity-Shake Flasks, Protein Folding and Drug Discovery. *Curr. Top. Med. Chem.* **2010**, *10*, 67–83.
- (534) Yap, C.; Xue, Y.; Li, H.; Li, Z.; Ung, C.; Han, L.; Zheng, C.; Cao, Z.; Chen, Y. Prediction of Compounds with Specific Pharmacodynamic, Pharmacokinetic or Toxicological Property by Statistical Learning Methods. *Mini-Rev. Med. Chem.* **2006**, *6*, 449–459.
- (535) Maltarollo, V. G.; Gertrudes, J. C.; Oliveira, P. R.; Honorio, K. M. Applying Machine Learning Techniques for ADME-Tox Prediction: a Review. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11*, 259–271.
- (536) Alqahtani, S. In Silico ADME-Tox Modeling: Progress and Prospects. *Expert Opin. Drug Metab. Toxicol.* **2017**, *13*, 1147–1158.
- (537) Lipinski, C. A. Lead-and Drug-Like Compounds: the Rule-of-Five Revolution. *Drug Discovery Today: Technol.* **2004**, *1*, 337–341.
- (538) Zhang, M.-Q.; Wilkinson, B. Drug Discovery Beyond the ‘Rule-of-Five’. *Curr. Opin. Biotechnol.* **2007**, *18*, 478–488.
- (539) Leeson, P. D.; Springthorpe, B. The Influence of Drug-Like Concepts on Decision-Making in Medicinal Chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890.
- (540) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space Beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21*, 1115–1142.
- (541) Hann, M. M.; Keserü, G. M. Finding the Sweet Spot: The Role of Nature and Nurture in Medicinal Chemistry. *Nat. Rev. Drug Discovery* **2012**, *11*, 355–365.
- (542) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discovery* **2015**, *14*, 111–129.
- (543) van de Waterbeemd, H. Improving Compound Quality Through in Vitro and in Silico Physicochemical Profiling. *Chem. Biodivers.* **2009**, *6*, 1760–1766.
- (544) Morphy, R.; Rankovic, Z. The Physicochemical Challenges of Designing Multiple Ligands. *J. Med. Chem.* **2006**, *49*, 4961–4970.
- (545) Wang, J.; Skolnik, S. Recent Advances in Physicochemical and ADMET Profiling in Drug Discovery. *Chem. Biodiversity* **2009**, *6*, 1887–1899.
- (546) Walther, B.; Vis, P.; Taylor, A. *Lipophilicity in Drug Action and Toxicology*; VCH Publishers: New York, 1996.
- (547) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin. Drug Discovery* **2010**, *5*, 235–248.
- (548) Arnott, J. A.; Planey, S. L. The Influence of Lipophilicity in Drug Discovery and Design. *Expert Opin. Drug Discovery* **2012**, *7*, 863–875.
- (549) Lin, S.-T.; Sandler, S. I. Prediction of Octanol–Water Partition Coefficients Using a Group Contribution Solvation Model. *Ind. Eng. Chem. Res.* **1999**, *38*, 4081–4091.
- (550) Nguyen, T.-B.; De Hemptinne, J.-C.; Creton, B.; Kontogeorgis, G. M. GC-PPC-SAFT Equation of State for VLE and LLE of Hydrocarbons and Oxygenated Compounds. Sensitivity Analysis. *Ind. Eng. Chem. Res.* **2013**, *52*, 7014–7029.
- (551) Chen, C.-S.; Lin, S.-T. Prediction of pH Effect on the Octanol–Water Partition Coefficient of Ionizable Pharmaceuticals. *Ind. Eng. Chem. Res.* **2016**, *55*, 9284–9294.
- (552) Bannan, C. C.; Calabró, G.; Kyu, D. Y.; Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/water. *J. Chem. Theory Comput.* **2016**, *12*, 4015–4024.
- (553) Yang, S. S.; Lu, W. C.; Gu, T. H.; Yan, L. M.; Li, G. Z. QSAR Study of n-Octanol/Water Partition Coefficient of Some Aromatic Compounds Using Support Vector Regression. *QSAR Comb. Sci.* **2009**, *28*, 175–182.
- (554) Zhang, H.; Xiang, M.-L.; Ma, C.-Y.; Huang, Q.; Li, W.; Xie, Y.; Wei, Y.-Q.; Yang, S.-Y. Three-Class Classification Models of LogS and LogP Derived by Using GA–CG–SVM Approach. *Mol. Diversity* **2009**, *13*, 261–268.
- (555) Bruneau, P.; McElroy, N. R. LogD 7.4 Modeling Using Bayesian Regularized Neural Networks. Assessment and Correction of the Errors of Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1379–1387.
- (556) Liao, Q.; Yao, J.; Yuan, S. SVM Approach for Predicting LogP. *Mol. Diversity* **2006**, *10*, 301–309.
- (557) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to Predict LogD Distribution Coefficient for Pfizer Proprietary Compounds. *J. Med. Chem.* **2004**, *47*, 5601–5604.
- (558) Zang, Q.; Mansouri, K.; Williams, A. J.; Judson, R. S.; Allen, D. G.; Casey, W. M.; Kleinstreuer, N. C. In Silico Prediction of

- Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *J. Chem. Inf. Model.* **2017**, *57*, 36–49.
- (559) Tetko, I. V. Associative Neural Network. *Neural Process. Lett.* **2002**, *16*, 187–199.
- (560) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.
- (561) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407–1421.
- (562) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol–Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47*, 2140–2148.
- (563) Remko, M. Acidity, Lipophilicity, Solubility, Absorption, and Polar Surface Area of Some ACE Inhibitors. *Chem. Pap.* **2007**, *61*, 133–141.
- (564) Kamiya, M.; Asanuma, D.; Kuranaga, E.; Takeishi, A.; Sakabe, M.; Miura, M.; Nagano, T.; Urano, Y.  $\beta$ -Galactosidase Fluorescence Probe with Improved Cellular Accumulation Based on a Spirocyclized Rhodol Scaffold. *J. Am. Chem. Soc.* **2011**, *133*, 12960–12963.
- (565) Asanuma, D.; Sakabe, M.; Kamiya, M.; Yamamoto, K.; Hiratake, J.; Ogawa, M.; Kosaka, N.; Choyke, P. L.; Nagano, T.; Kobayashi, H.; et al. Sensitive  $\beta$ -Galactosidase-Targeting Fluorescence Probe for Visualizing Small Peritoneal Metastatic Tumours in Vivo. *Nat. Commun.* **2015**, *6*, 6463.
- (566) Wilson, J. J.; Lippard, S. J. In Vitro Anticancer Activity of cis-Diammineplatinum (II) Complexes with  $\beta$ -Diketonate Leaving Group Ligands. *J. Med. Chem.* **2012**, *55*, 5326–5336.
- (567) Bahmani, A.; Saaidpour, S.; Rostami, A. A Simple, Robust and Efficient Computational Method for n-Octanol/Water Partition Coefficients of Substituted Aromatic Drugs. *Sci. Rep.* **2017**, *7*, 5760.
- (568) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046–1051.
- (569) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.
- (570) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726–741.
- (571) Zhang, J.; Tuguldur, B.; van der Spoel, D. Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation. *J. Chem. Inf. Model.* **2015**, *55*, 1192–1201.
- (572) Fuchs, J.-A.; Grisoni, F.; Kossenjans, M.; Hiss, J. A.; Schneider, G. Lipophilicity Prediction of Peptides and Peptide Derivatives by Consensus Machine Learning. *MedChemComm* **2018**, *9*, 1538–1546.
- (573) Stegemann, S.; Leveiller, F.; Franchi, D.; De Jong, H.; Lindén, H. When Poor Solubility Becomes an Issue: From Early Stage to Proof of Concept. *Eur. J. Pharm. Sci.* **2007**, *31*, 249–261.
- (574) McDonagh, J. L. N.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **2014**, *54*, 844–856.
- (575) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155–1158.
- (576) Dearden, J. C. In Silico Prediction of Aqueous Solubility. *Expert Opin. Drug Discovery* **2006**, *1*, 31–52.
- (577) Delaney, J. S. Predicting Aqueous Solubility from Structure. *Drug Discovery Today* **2005**, *10*, 289–295.
- (578) Johnson, S. R.; Zheng, W. Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption. *AAPS J.* **2006**, *8*, E27–E40.
- (579) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (580) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355–366.
- (581) Schneider, G. Prediction of Drug-Like Properties. In *Madame Curie Bioscience Database 2000–2013*; Landes Bioscience: Austin, TX, 2013.
- (582) Skyner, R.; McDonagh, J.; Groom, C.; Van Mourik, T.; Mitchell, J. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191.
- (583) Wang, J.; Hou, T. Recent Advances on Aqueous Solubility Prediction. *Comb. Chem. High Throughput Screening* **2011**, *14*, 328–338.
- (584) Jain, N.; Yalkowsky, S. H. Estimation of the Aqueous Solubility I: Application to Organic Nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234–252.
- (585) Ran, Y.; He, Y.; Yang, G.; Johnson, J. L.; Yalkowsky, S. H. Estimation of Aqueous Solubility of Organic Compounds by Using the General Solubility Equation. *Chemosphere* **2002**, *48*, 487–509.
- (586) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
- (587) Wang, J.; Hou, T.; Xu, X. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571–581.
- (588) Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. *J. Chem. Inf. Model.* **2012**, *52*, 420–428.
- (589) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Model.* **1992**, *32*, 474–482.
- (590) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061–2077.
- (591) Timmerman, H.; Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2002.
- (592) Fühner, H. The Aqueous Solubility of Homologous Series. *Ber. Dtsch. Chem. Ges. B* **1924**, *57B*, 510–515.
- (593) Kamlet, M. J.; Doherty, R. M.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. Linear Solvation Energy Relationships: 36. Molecular Properties Governing Solubilities of Organic Nonelectrolytes in Water. *J. Pharm. Sci.* **1986**, *75*, 338–349.
- (594) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Design*; Academic Press: New York, 1976.
- (595) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Letchworth, Herts, England, 1986.
- (596) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- (597) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (598) Zhou, D.; Alelyunas, Y.; Liu, R. Scores of Extended Connectivity Fingerprint as Descriptors in QSPR Study of Melting Point and Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *48*, 981–987.
- (599) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Binary Classification of Aqueous Solubility Using Support Vector Machines with Reduction and Recombination Feature Selection. *J. Chem. Inf. Model.* **2011**, *51*, 229–236.
- (600) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
- (601) 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**, *S3*, 1563–1575.

- (602) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.
- (603) Hewitt, M.; Cronin, M. T.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49*, 2572–2587.
- (604) Llinas, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.
- (605) Palmer, D. S.; Mitchell, J. B. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? *Mol. Pharmaceutics* **2014**, *11*, 2962–2972.
- (606) Hopfinger, A. J.; Esposito, E. X.; Llinas, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2009**, *49*, 1–5.
- (607) Suenderhauf, C.; Hammann, F.; Maunz, A.; Helma, C.; Huwyler, J. r. Combinatorial QSAR Modeling of Human Intestinal Absorption. *Mol. Pharmaceutics* **2011**, *8*, 213–224.
- (608) Zheng, T.; Hopfinger, A. J.; Esposito, E. X.; Liu, J.; Tseng, Y. J. Membrane-Interaction Quantitative Structure–Activity Relationship (MI-QSAR) Analyses of Skin Penetration Enhancers. *J. Chem. Inf. Model.* **2008**, *48*, 1238–1256.
- (609) Seelig, A. The Role of Size and Charge for Blood–Brain Barrier Permeation of Drugs and Fatty Acids. *J. Mol. Neurosci.* **2007**, *33*, 32–41.
- (610) Leung, S. S.; Mijalkovic, J.; Borrelli, K.; Jacobson, M. P. Testing Physical Models of Passive Membrane Permeation. *J. Chem. Inf. Model.* **2012**, *52*, 1621–1636.
- (611) Rezai, T.; Yu, B.; Millhauser, G. L.; Jacobson, M. P.; Lokey, R. S. Testing the Conformational Hypothesis of Passive Membrane Permeability Using Synthetic Cyclic Peptide Diastereomers. *J. Am. Chem. Soc.* **2006**, *128*, 2510–2511.
- (612) Rezai, T.; Bock, J. E.; Zhou, M. V.; Kalyanaraman, C.; Lokey, R. S.; Jacobson, M. P. Conformational Flexibility, Internal Hydrogen Bonding, and Passive Membrane Permeability: Successful in Silico Prediction of the Relative Permeabilities of Cyclic Peptides. *J. Am. Chem. Soc.* **2006**, *128*, 14073–14080.
- (613) Rand, A. C.; Leung, S. S.; Rotter, C. J.; Sharma, R.; Kalgutkar, A. S.; Zhang, Y.; Varma, M. V.; Farley, K. A.; Khunte, B.; Limberakis, C.; et al. Optimizing PK Properties of Cyclic Peptides: The Effect of Side Chain Substitutions on Permeability and Clearance. *MedChem-Comm* **2012**, *3*, 1282–1289.
- (614) Tang, F.; Horie, K.; Borchardt, R. T. Are MDCK Cells Transfected with the Human MDR1 Gene a Good Model of the Human Intestinal Mucosa? *Pharm. Res.* **2002**, *19*, 765–772.
- (615) Broccatelli, F.; Salphati, L.; Plise, E.; Cheong, J.; Gobbi, A.; Lee, M.-L.; Aliagas, I. Predicting Passive Permeability of Drug-Like Molecules from Chemical Structure: Where Are We? *Mol. Pharmaceutics* **2016**, *13*, 4199–4208.
- (616) Ballard, P.; Brassil, P.; Bui, K. H.; Dolgos, H.; Petersson, C.; Tunek, A.; Webborn, P. J. The Right Compound in the Right Assay at the Right Time: An Integrated Discovery DMPK Strategy. *Drug Metab. Rev.* **2012**, *44*, 224–252.
- (617) Hubatsch, I.; Ragnarsson, E. G.; Artursson, P. Determination of Drug Permeability and Prediction of Drug Absorption in Caco-2 Monolayers. *Nat. Protoc.* **2007**, *2*, 2111–2119.
- (618) Ungell, A.-L. B. Caco-2 Replace or Refine? *Drug Discovery Today: Technol.* **2004**, *1*, 423–430.
- (619) Angelis, I. D.; Turco, L. Caco-2 Cells as a Model for Intestinal Absorption. *Curr. Protoc. Toxicol.* **2011**, *47*, 1.
- (620) Fredlund, L.; Winiwarter, S.; Hilgendorf, C. In Vitro Intrinsic Permeability: A Transporter-Independent Measure of Caco-2 Cell Permeability in Drug Design and Development. *Mol. Pharmaceutics* **2017**, *14*, 1601–1609.
- (621) Dixon, S. L.; Duan, J.; Smith, E.; Von Bargen, C. D.; Sherman, W.; Repasky, M. P. AutoQSAR: An Automated Machine Learning Tool for Best-Practice Quantitative Structure–Activity Relationship Modeling. *Future Med. Chem.* **2016**, *8*, 1825–1839.
- (622) Avdeef, A.; Artursson, P.; Neuhoff, S.; Lazorova, L.; Gråsjö, J.; Tavelin, S. Caco-2 Permeability of Weakly Basic Drugs Predicted with the Double-Sink PAMPA pKaflux Method. *Eur. J. Pharm. Sci.* **2005**, *24*, 333–349.
- (623) Avdeef, A.; Bendels, S.; Di, L.; Faller, B.; Kansy, M.; Sugano, K.; Yamauchi, Y. PAMPA—Critical Factors for Better Predictions of Absorption. *J. Pharm. Sci.* **2007**, *96*, 2893–2909.
- (624) Avdeef, A. The Rise of PAMPA. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 325–342.
- (625) Sun, H.; Nguyen, K.; Kerns, E.; Yan, Z.; Yu, K. R.; Shah, P.; Jadhav, A.; Xu, X. Highly Predictive and Interpretable Models for PAMPA Permeability. *Bioorg. Med. Chem.* **2017**, *25*, 1266–1276.
- (626) Zhang, J. h.; Liu, Z. m.; Liu, W. r. QSPR Study for Prediction of Boiling Points of 2475 Organic Compounds Using Stochastic Gradient Boosting. *J. Chemom.* **2014**, *28*, 161–167.
- (627) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. Why Are Some Properties More Difficult to Predict than Others? A Study of QSPR Models of Solubility, Melting point, and LogP. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.
- (628) Bhattacharai, B.; Teetz, W.; Liu, T.; Öberg, T.; Jeliazkova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E.; et al. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Mol. Inf.* **2011**, *30*, 189–204.
- (629) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminform.* **2018**, *10*, 10.
- (630) Rupp, M.; Körner, R.; Tetko, I. V. Estimation of Acid Dissociation Constants Using Graph Kernels. *Mol. Inf.* **2010**, *29*, 731–740.
- (631) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457–472.
- (632) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Gancia, E.; Manallack, D. T. Estimation of pKa Using Semiempirical Molecular Orbital Methods. Part 2: Application to Amines, Anilines and Various Nitrogen Containing Heterocyclic Compounds. *Quant. Struct.-Act. Relat.* **2002**, *21*, 473–485.
- (633) Zhou, T.; Jhamb, S.; Liang, X.; Sundmacher, K.; Gani, R. Prediction of Acid Dissociation Constants of Organic Compounds Using Group Contribution Methods. *Chem. Eng. Sci.* **2018**, *183*, 95–105.
- (634) Li, M.; Zhang, H.; Chen, B.; Wu, Y.; Guan, L. Prediction of pKa Values for Neutral and Basic Drugs based on Hybrid Artificial Intelligence Methods. *Sci. Rep.* **2018**, *8*, 3991.
- (635) Rankovic, Z. CNS Drug Design: Balancing Physicochemical Properties for Optimal Brain Exposure. *J. Med. Chem.* **2015**, *58*, 2584–2608.
- (636) Kaitin, K. Obstacles and Opportunities in New Drug Development. *Clin. Pharmacol. Ther.* **2008**, *83*, 210–212.
- (637) Ghosh, J.; Lawless, M. S.; Waldman, M.; Gombar, V.; Fraczkiewicz, R. Modeling ADMET. In *In Silico Methods for Predicting Drug Toxicity*; Humana Press: New York, 2016.
- (638) Cox, P. B.; Gregg, R. J.; Vasudevan, A. Abbott Physicochemical Tiering (APT)—A Unified Approach to HTS Triage. *Bioorg. Med. Chem.* **2012**, *20*, 4564–4573.
- (639) Stoner, C. L.; Troutman, M.; Gao, H.; Johnson, K.; Stankovic, C.; Brodfuehrer, J.; Gifford, E.; Chang, M. Moving in Silico Screening into Practice: A Minimalist Approach to Guide Permeability Screening!! *Lett. Drug Des. Discovery* **2006**, *3*, 575–581.
- (640) Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.

- (641) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME Properties with Substructure Pattern Recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034–1041.
- (642) Hou, T.; Li, Y.; Zhang, W.; Wang, J. Recent Developments of in Silico Predictions of Intestinal Absorption and Oral Bioavailability. *Comb. Chem. High Throughput Screening* **2009**, *12*, 497–506.
- (643) Kuentz, M. Prediction of Drug Absorption: Different Modeling Approaches from Discovery to Clinical Development. *Expert Rev. Clin. Pharmacol.* **2009**, *2*, 217–219.
- (644) Talevi, A.; Goodarzi, M.; Ortiz, E. V.; Duchowicz, P. R.; Bellera, C. L.; Pesce, G.; Castro, E. A.; Bruno-Blanch, L. E. Prediction of Drug Intestinal Absorption by New Linear and Non-linear QSPR. *Eur. J. Med. Chem.* **2011**, *46*, 218–228.
- (645) Billat, P.-A.; Roger, E.; Faure, S.; Lagarce, F. Models for Drug Absorption from the Small Intestine: Where Are We and Where Are We Going? *Drug Discovery Today* **2017**, *22*, 761–775.
- (646) Kumar, R.; Sharma, A.; Siddiqui, M. H.; Tiwari, R. K. Prediction of Human Intestinal Absorption of Compounds Using Artificial Intelligence Techniques. *Curr. Drug Discovery Technol.* **2017**, *14*, 244–254.
- (647) Kumar, R.; Sharma, A.; Siddiqui, M. H.; Tiwari, R. K. Promises of Machine Learning Approaches in Prediction of Absorption of Compounds. *Mini Rev. Med. Chem.* **2018**, *18*, 196–207.
- (648) Wang, N.-N.; Huang, C.; Dong, J.; Yao, Z.-J.; Zhu, M.-F.; Deng, Z.-K.; Lv, B.; Lu, A.-P.; Chen, A. F.; Cao, D.-S. Predicting Human Intestinal Absorption with Modified Random Forest Approach: A Comprehensive Evaluation of Molecular Representation, Unbalanced Data, and Applicability Domain Issues. *RSC Adv.* **2017**, *7*, 19007–19018.
- (649) Basant, N.; Gupta, S.; Singh, K. P. Predicting Human Intestinal Absorption of Diverse Chemicals Using Ensemble Learning Based QSAR Modeling Approaches. *Comput. Biol. Chem.* **2016**, *61*, 178–196.
- (650) Shin, M.; Jang, D.; Nam, H.; Lee, K. H.; Lee, D. Predicting the Absorption Potential of Chemical Compounds through a Deep Learning Approach. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2018**, *15*, 432–440.
- (651) Wang, N.-N.; Dong, J.; Deng, Y.-H.; Zhu, M.-F.; Wen, M.; Yao, Z.-J.; Lu, A.-P.; Wang, J.-B.; Cao, D.-S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56*, 763–773.
- (652) Lambrinidis, G.; Vallianatou, T.; Tsantili-Kakoulidou, A. In Vitro, in Silico and Integrated Strategies for the Estimation of Plasma Protein Binding. A Review. *Adv. Drug Delivery Rev.* **2015**, *86*, 27–45.
- (653) Zhivkova, Z. Studies on Drug–Human Serum Albumin Binding: The Current State of the Matter. *Curr. Pharm. Des.* **2015**, *21*, 1817–1830.
- (654) Kopecký, V. r.; Ettrich, R.; Hofbauerová, K.; Baumruk, V. Structure of Human  $\alpha$ 1-Acid Glycoprotein and Its High-Affinity Binding Site. *Biochem. Biophys. Res. Commun.* **2003**, *300*, 41–46.
- (655) Kuchinskiene, Z.; Carlson, L. Composition, Concentration, and Size of Low Density Lipoproteins and of Subfractions of Very Low Density Lipoproteins from Serum of Normal Men and Women. *J. Lipid Res.* **1982**, *23*, 762–769.
- (656) Schmidt, S.; Gonzalez, D.; Derendorf, H. Significance of Protein Binding in Pharmacokinetics and Pharmacodynamics. *J. Pharm. Sci.* **2010**, *99*, 1107–1122.
- (657) Jambhekar, S. S. Physicochemical and Biopharmaceutical Properties of Drug Substances and Pharmacokinetics. In *Foye's Principles of Medicinal Chemistry*; Lippincott Williams and Wilkins: Philadelphia, 2003.
- (658) Hochman, J.; Tang, C.; Prueksaritanont, T. Drug–Drug Interactions Related to Altered Absorption and Plasma Protein Binding: Theoretical and Regulatory Considerations, and an Industry Perspective. *J. Pharm. Sci.* **2015**, *104*, 916–929.
- (659) Ma, C.-Y.; Yang, S.-Y.; Zhang, H.; Xiang, M.-L.; Huang, Q.; Wei, Y.-Q. Prediction Models of Human Plasma Protein Binding Rate and Oral Bioavailability Derived by Using GA–CG–SVM Method. *J. Pharm. Biomed. Anal.* **2008**, *47*, 677–682.
- (660) Gleeson, M. P. Plasma Protein Binding Affinity and Its Relationship to Molecular Structure: an in-Silico Analysis. *J. Med. Chem.* **2007**, *50*, 101–112.
- (661) Ghafourian, T.; Amin, Z. QSAR Models for the Prediction of Plasma Protein Binding. *BioImpacts* **2013**, *3*, 21–27.
- (662) Chen, L.; Chen, X. Results of Molecular Docking as Descriptors to Predict Human Serum Albumin Binding Affinity. *J. Mol. Graphics Modell.* **2012**, *33*, 35–43.
- (663) Sun, L.; Yang, H.; Li, J.; Wang, T.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Compounds Binding to Human Plasma Proteins by QSAR Models. *ChemMedChem* **2018**, *13*, 572–581.
- (664) Basant, N.; Gupta, S.; Singh, K. Predicting Binding Affinities of Diverse Pharmaceutical Chemicals to Human Serum Plasma Proteins Using QSPR Modelling Approaches. *SAR QSAR Environ. Res.* **2016**, *27*, 67–85.
- (665) Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R. Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J. Chem. Inf. Model.* **2016**, *56*, 2243–2252.
- (666) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure–Information Representation. *J. Med. Chem.* **2006**, *49*, 7169–7181.
- (667) Moda, T. L.; Torres, L. G.; Carrara, A. E.; Andricopulo, A. D. PK/DB: Database for Pharmacokinetic Properties and Predictive in silico ADME Models. *Bioinformatics* **2008**, *24*, 2270–2271.
- (668) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
- (669) Suenderhauf, C.; Hammann, F.; Huwyler, J. Computational Prediction of Blood–Brain Barrier Permeability Using Decision Tree Induction. *Molecules* **2012**, *17*, 10429–10445.
- (670) Wang, Z.; Yan, A.; Yuan, Q. Classification of Blood–Brain Barrier Permeation by Kohonen’s Self-Organizing Neural Network (KohNN) and Support Vector Machine (SVM). *QSAR Comb. Sci.* **2009**, *28*, 989–994.
- (671) Deconinck, E.; Zhang, M.; Petitet, F.; Dubus, E.; Ijjaali, I.; Coomans, D.; Vander Heyden, Y. Boosted Regression Trees, Multivariate Adaptive Regression Splines and Their Two-Step Combinations with Multiple Linear Regression or Partial Least Squares to Predict Blood–Brain Barrier Passage: A Case Study. *Anal. Chim. Acta* **2008**, *609*, 13–23.
- (672) Bujak, R.; Struck-Lewicka, W.; Kaliszan, M.; Kaliszan, R.; Markuszewski, M. J. Blood–Brain Barrier Permeability Mechanisms in View of Quantitative Structure–Activity Relationships (QSAR). *J. Pharm. Biomed. Anal.* **2015**, *108*, 29–37.
- (673) Carpenter, T. S.; Kirshner, D. A.; Lau, E. Y.; Wong, S. E.; Nilmeier, J. P.; Lightstone, F. C. A Method to Predict Blood–Brain Barrier Permeability of Drug-Like Compounds Using Molecular Dynamics Simulations. *Biophys. J.* **2014**, *107*, 630–641.
- (674) Kumar, R.; Sharma, A.; Tiwari, R. K. Can We Predict Blood Brain Barrier Permeability of Ligands Using Computational Approaches? *Interdiscip. Sci.: Comput. Life Sci.* **2013**, *5*, 95–101.
- (675) Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and Prediction of a Large Blood–Brain Distribution Data Set—an LFER Study. *Eur. J. Med. Chem.* **2001**, *36*, 719–730.
- (676) Ghose, A. K.; Herbertz, T.; Hudkins, R. L.; Dorsey, B. D.; Mallamo, J. P. Knowledge-Based, Central Nervous System (CNS) Lead Selection and Lead Optimization for CNS Drug Discovery. *ACS Chem. Neurosci.* **2012**, *3*, 50–68.
- (677) Chen, Y.; Zhu, Q.-J.; Pan, J.; Yang, Y.; Wu, X.-P. A Prediction Model for Blood–Brain Barrier Permeation and Analysis on Its Parameter Biologically. *Comput. Methods Programs Biomed.* **2009**, *95*, 280–287.

- (678) Subramanian, G.; Kitchen, D. B. Computational Models to Predict Blood–Brain Barrier Permeation and CNS Activity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 643–664.
- (679) Pardridge, W. M. The Blood–Brain Barrier: Bottleneck in Brain Drug Development. *NeuroRx* **2005**, *2*, 3–14.
- (680) Norinder, U.; Haeberlein, M. Computational Approaches to the Prediction of the Blood–Brain Distribution. *Adv. Drug Delivery Rev.* **2002**, *54*, 291–313.
- (681) Gao, Z.; Chen, Y.; Cai, X.; Xu, R. Predict Drug Permeability to Blood–Brain-Barrier from Clinical Phenotypes: Drug Side Effects and Drug Indications. *Bioinformatics* **2016**, *33*, 901–908.
- (682) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A Side Effect Resource to Capture Phenotypic Effects of Drugs. *Mol. Syst. Biol.* **2010**, *6*, 343.
- (683) Kell, D. B.; Goodacre, R. Metabolomics and Systems Pharmacology: Why and How to Model the Human Metabolic Network for Drug Discovery. *Drug Discovery Today* **2014**, *19*, 171–182.
- (684) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617–648.
- (685) De Graaf, C.; Vermeulen, N. P.; Feenstra, K. A. Cytochrome P450 in Silico: An Integrative Modeling Approach. *J. Med. Chem.* **2005**, *48*, 2725–2755.
- (686) Crivori, P.; Poggesi, I. Computational Approaches for Predicting CYP-Related Metabolism Properties in the Screening of New Drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795–808.
- (687) Fox, T.; Kriegl, J. M. Machine Learning Techniques for in Silico Modeling of Drug Metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579–1591.
- (688) Cerny, M. A. Prevalence of Non–Cytochrome P450–Mediated Metabolism in Food and Drug Administration–Approved Oral and Intravenous Drugs: 2006–2015. *Drug Metab. Dispos.* **2016**, *44*, 1246–1252.
- (689) Kwon, Y. *Handbook of Essential Pharmacokinetics, Pharmacodynamics and Drug Metabolism for Industrial Scientists*; Kluwer Academic/ Plenum Publishers: New York, 2001.
- (690) Dixit, V. A.; Lal, L. A.; Agrawal, S. R. Recent Advances in the Prediction of Non-CYP450-Mediated Drug Metabolism. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2017**, *7*, e1323.
- (691) Moroy, G.; Martiny, V. Y.; Vayer, P.; Villoutreix, B. O.; Miteva, M. A. Toward in Silico Structure-Based ADMET Prediction in Drug Discovery. *Drug Discovery Today* **2012**, *17*, 44–55.
- (692) Sun, H.; Scott, D. O. Structure-Based Drug Metabolism Predictions for Drug Design. *Chem. Biol. Drug Des.* **2010**, *75*, 3–17.
- (693) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting Drug Metabolism: Experiment and/or Computation? *Nat. Rev. Drug Discovery* **2015**, *14*, 387–404.
- (694) Vasanthanathan, P.; Tabouret, O.; Oostenbrink, C.; Vermeulen, N. P.; Olsen, L.; Jørgensen, F. S. Classification of Cytochrome P450 1A2 Inhibitors and Non-Inhibitors by Machine Learning Techniques. *Drug Metab. Dispos.* **2009**, *37*, 658–664.
- (695) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-w.; Bennett, K. P.; Breneman, C. M. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J. Chem. Inf. Model.* **2011**, *51*, 1667–1689.
- (696) Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M. RS-Predictor Models Augmented with SMARTCyp Reactivities: Robust Metabolic Regioselectivity Predictions for Nine CYP Isozymes. *J. Chem. Inf. Model.* **2012**, *52*, 1637–1659.
- (697) Zaretzki, J.; Bergeron, C.; Huang, T.-w.; Rydberg, P.; Swamidass, S. J.; Breneman, C. M. RS-WebPredictor: A Server for Predicting CYP-Mediated Sites of Metabolism on Drug-Like Molecules. *Bioinformatics* **2013**, *29*, 497–498.
- (698) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96–100.
- (699) Finkelman, A. R.; Göller, A. H.; Schneider, G. Site of Metabolism Prediction Based on ab Initio Derived Atom Representations. *ChemMedChem* **2017**, *12*, 606–612.
- (700) Zaretzki, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *J. Chem. Inf. Model.* **2013**, *53*, 3373–3383.
- (701) Matlock, M. K.; Hughes, T. B.; Swamidass, S. J. XenoSite Server: A Web-Available Site of Metabolism Prediction Tool. *Bioinformatics* **2015**, *31*, 1136–1137.
- (702) Kirchmair, J.; Williamson, M. J.; Afzal, A. M.; Tyzack, J. D.; Choy, A. P.; Howlett, A.; Rydberg, P.; Glen, R. C. FASt MEtabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *J. Chem. Inf. Model.* **2013**, *53*, 2896–2907.
- (703) Šicho, M.; De Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *J. Chem. Inf. Model.* **2017**, *57*, 1832–1846.
- (704) Dang, N. L.; Hughes, T. B.; Krishnamurthy, V.; Swamidass, S. J. A Simple Model Predicts UGT-Mediated Metabolism. *Bioinformatics* **2016**, *32*, 3183–3189.
- (705) Zaretzki, J.; Boehm, K. M.; Swamidass, S. J. Improved Prediction of CYP-Mediated Metabolism with Chemical Fingerprints. *J. Chem. Inf. Model.* **2015**, *55*, 972–982.
- (706) Zaretzki, J. M.; Browning, M. R.; Hughes, T. B.; Swamidass, S. J. Extending P450 Site-of-Metabolism Models with Region-Resolution Data. *Bioinformatics* **2015**, *31*, 1966–1973.
- (707) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Modeling Epoxidation of Drug-Like Molecules with a Deep Machine Learning Network. *ACS Cent. Sci.* **2015**, *1*, 168–180.
- (708) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Site of Reactivity Models Predict Molecular Reactivity of Diverse Chemicals with Glutathione. *Chem. Res. Toxicol.* **2015**, *28*, 797–809.
- (709) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. *ACS Cent. Sci.* **2016**, *2*, 529–537.
- (710) Hughes, T. B.; Swamidass, S. J. Deep Learning to Predict the Formation of Quinone Species in Drug Metabolism. *Chem. Res. Toxicol.* **2017**, *30*, 642–656.
- (711) Dixit, V. A.; Deshpande, S. Advances in Computational Prediction of Regioselective and Isoform-Specific Drug Metabolism Catalyzed by CYP450s. *ChemistrySelect* **2016**, *1*, 6571–6597.
- (712) Kumar, R.; Sharma, A.; Haris Siddiqui, M.; Kumar Tiwari, R. Prediction of Metabolism of Drugs Using Artificial Intelligence: How Far Have We Reached? *Curr. Drug Metab.* **2016**, *17*, 129–141.
- (713) Lewis, D. F. On the Recognition of Mammalian Microsomal Cytochrome P450 Substrates and Their Characteristics: Towards the Prediction of Human p450 Substrate Specificity and Metabolism. *Biochem. Pharmacol.* **2000**, *60*, 293–306.
- (714) Shao, C.-Y.; Su, B.-H.; Tu, Y.-S.; Lin, C.; Lin, O. A.; Tseng, Y. J. CypRules: A Rule-Based P450 Inhibition Prediction Server. *Bioinformatics* **2015**, *31*, 1869–1871.
- (715) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* **2005**, *48*, 6970–6979.
- (716) Cruciani, G.; Valeri, A.; Goracci, L.; Pellegrino, R. M.; Buonerba, F.; Baroni, M. Flavin Monooxygenase Metabolism: Why Medicinal Chemists Should Matter. *J. Med. Chem.* **2014**, *57*, 6183–6196.
- (717) Mishra, N. K.; Agarwal, S.; Raghava, G. P. Prediction of Cytochrome P450 Isoform Responsible for Metabolizing a Drug Molecule. *BMC Pharmacol.* **2010**, *10*, 8.
- (718) Rydberg, P.; Olsen, L. Ligand-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *ACS Med. Chem. Lett.* **2012**, *3*, 69–73.

- (719) Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: Prediction of Cytochromes P450 Inhibition. *Bioinformatics* **2013**, *29*, 2051–2052.
- (720) Lombardo, F.; Jing, Y. In Silico Prediction of Volume of Distribution in Humans. Extensive Data Set and the Exploration of Linear and Nonlinear Methods Coupled with Molecular Interaction Fields Descriptors. *J. Chem. Inf. Model.* **2016**, *56*, 2042–2052.
- (721) Berellini, G.; Waters, N. J.; Lombardo, F. In Silico Prediction of Total Human Plasma Clearance. *J. Chem. Inf. Model.* **2012**, *52*, 2069–2078.
- (722) Toshimoto, K.; Wakayama, N.; Kusama, M.; Maeda, K.; Sugiyama, Y.; Akiyama, Y. In Silico Prediction of Major Drug Clearance Pathways by Support Vector Machines with Feature-Selected Descriptors. *Drug Metab. Dispos.* **2014**, *42*, 1811–1819.
- (723) Cañada, A.; Capella-Gutierrez, S.; Rabal, O.; Oyarzabal, J.; Valencia, A.; Krallinger, M. LimTox: A Web Tool for Applied Text Mining of Adverse Event and Toxicity Associations of Compounds, Drugs and Genes. *Nucleic Acids Res.* **2017**, *45*, W484–W489.
- (724) Pires, D. E.; Blundell, T. L.; Ascher, D. B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* **2015**, *58*, 4066–4072.
- (725) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. AdmetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099–3105.
- (726) Patlewicz, G.; Jeliazkova, N.; Safford, R.; Worth, A.; Aleksiev, B. An Evaluation of the Implementation of the Cramer Classification Scheme in the Toxtree Software. *SAR QSAR Environ. Res.* **2008**, *19*, 495–524.
- (727) Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf. Model.* **2017**, *57*, 2672–2685.
- (728) Chavan, S.; Abdelaziz, A.; Wiklander, J. G.; Nicholls, I. A. A k-Nearest Neighbor Classification of hERG K<sup>+</sup> Channel Blockers. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 229–236.
- (729) Chavan, S.; Friedman, R.; Nicholls, I. A. Acute Toxicity-Supported Chronic Toxicity Prediction: A k-Nearest Neighbor Coupled Read-Across Strategy. *Int. J. Mol. Sci.* **2015**, *16*, 11659–11677.
- (730) Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Drug Induced Liver Toxicity using Substructure Pattern Recognition Method. *Mol. Inf.* **2016**, *35*, 136–144.
- (731) Enslein, K. A. Toxicity Estimation Model. *J. Environ. Pathol. Toxicol.* **1978**, *2*, 115–121.
- (732) Enslein, K.; Lander, T.; Tomb, M.; Craig, P. A Predictive Model for Estimating Rat Oral LD<sub>50</sub> Values. *Toxicol. Ind. Health* **1989**, *5*, 261–387.
- (733) Uesawa, Y. Rigorous Selection of Random Forest Models for Identifying Compounds that Activate Toxicity-Related Pathways. *Frontiers in Environmental Science* **2016**, *4*, 9.
- (734) Zhu, X.-W.; Xin, Y.-J.; Chen, Q.-H. Chemical and In Vitro Biological Information to Predict Mouse Liver Toxicity Using Recursive Random Forests. *SAR QSAR Environ. Res.* **2016**, *27*, 559–572.
- (735) Zhang, H.; Ren, J.-X.; Kang, Y.-L.; Bo, P.; Liang, J.-Y.; Ding, L.; Kong, W.-B.; Zhang, J. Development of Novel in Silico Model for Developmental Toxicity Assessment by Using Naïve Bayes Classifier Method. *Reprod. Toxicol.* **2017**, *71*, 8–15.
- (736) Zhang, H.; Yu, P.; Ren, J.-X.; Li, X.-B.; Wang, H.-L.; Ding, L.; Kong, W.-B. Development of Novel Prediction Model for Drug-Induced Mitochondrial Toxicity by Using Naïve Bayes Classifier Method. *Food Chem. Toxicol.* **2017**, *110*, 122–129.
- (737) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (738) Takigawa, I.; Mamitsuka, H. Graph Mining: Procedure, Application to Drug Discovery and Recent Advances. *Drug Discovery Today* **2013**, *18*, 50–57.
- (739) Rupp, M.; Schneider, G. Graph Kernels for Molecular Similarity. *Mol. Inf.* **2010**, *29*, 266–273.
- (740) Lei, T.; Li, Y.; Song, Y.; Li, D.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery: 15. Accurate Prediction of Rat Oral Acute Toxicity Using Relevance Vector Machine and Consensus Modeling. *J. Cheminform.* **2016**, *8*, 6.
- (741) Lei, T.; Chen, F.; Liu, H.; Sun, H.; Kang, Y.; Li, D.; Li, Y.; Hou, T. ADMET Evaluation in Drug Discovery. Part 17: Development of Quantitative and Qualitative Prediction Models for Chemical-Induced Respiratory Toxicity. *Mol. Pharmaceutics* **2017**, *14*, 2407–2421.
- (742) Lei, T.; Sun, H.; Kang, Y.; Zhu, F.; Liu, H.; Zhou, W.; Wang, Z.; Li, D.; Li, Y.; Hou, T. ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning Approaches. *Mol. Pharmaceutics* **2017**, *14*, 3935–3953.
- (743) Tipping, M. E. The Relevance Vector Machine. In *Advances in Neural Information Processing Systems*, 2000, 652–658.
- (744) Pandya, R.; Pandya, J. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *Int. J. Comput. Appl.* **2015**, *117*, 18–21.
- (745) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y. *Xgboost: Extreme Gradient Boosting*. R package version 0.4–2, 2015.
- (746) Tharwat, A.; Moemen, Y. S.; Hassanien, A. E. A Predictive Model for Toxicity Effects Assessment of Biotransformed Hepatic Drugs Using Iterative Sampling Method. *Sci. Rep.* **2016**, *6*, 38660.
- (747) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068–2076.
- (748) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55*, 2085–2093.
- (749) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 125–134.
- (750) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
- (751) Koch, U.; Hamacher, M.; Nussbaumer, P. Cheminformatics at the Interface of Medicinal Chemistry and Proteomics. *Biochim. Biophys. Acta, Proteins Proteomics* **2014**, *1844*, 156–161.
- (752) Glicksberg, B. S.; Li, L.; Cheng, W.-Y.; Shameer, K.; Hakenberg, J.; Castellanos, R.; Ma, M.; Shi, L.; Shah, H.; Dudley, J. T. An Integrative Pipeline for Multi-Modal Discovery of Disease Relationships. In *Pacific Symposium on Biocomputing 2015*; World Scientific Publishing Co Pte Ltd, 2014; pp 407–418.
- (753) Piro, R. M. Network Medicine: Linking Disorders. *Hum. Genet.* **2012**, *131*, 1811–1820.
- (754) Sardana, D.; Zhu, C.; Zhang, M.; Gudivada, R. C.; Yang, L.; Jegga, A. G. Drug Repositioning for Orphan Diseases. *Briefings Bioinf.* **2011**, *12*, 346–356.
- (755) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673–683.
- (756) Dudley, J. T.; Deshpande, T.; Butte, A. J. Exploiting Drug-Disease Relationships for Computational Drug Repositioning. *Briefings Bioinf.* **2011**, *12*, 303–311.
- (757) Hurle, M.; Yang, L.; Xie, Q.; Rajpal, D.; Sanseau, P.; Agarwal, P. Computational Drug Repositioning: from Data to Therapeutics. *Clin. Pharmacol. Ther.* **2013**, *93*, 335–341.
- (758) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; et al. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.
- (759) Haupt, V. J.; Schroeder, M. Old Friends in New Guise: Repositioning of Known Drugs with Structural Bioinformatics. *Briefings Bioinf.* **2011**, *12*, 312–326.
- (760) De Franchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of Protein Kinase Inhibitors to SynapsinI

- Inferred from Pair-Wise Binding Site Similarity Measurements. *PLoS One* **2010**, *5*, e12214.
- (761) Kinnings, S. L.; Liu, N.; Buchmeier, N.; Tonge, P. J.; Xie, L.; Bourne, P. E. Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *PLoS Comput. Biol.* **2009**, *5*, e1000423.
- (762) Butcher, E. C.; Berg, E. L.; Kunkel, E. J. Systems Biology in Drug Discovery. *Nat. Biotechnol.* **2004**, *22*, 1253–1259.
- (763) Lamb, J.; Crawford, E. D.; Peck, D.; Modell, J. W.; Blat, I. C.; Wrobel, M. J.; Lerner, J.; Brunet, J.-P.; Subramanian, A.; Ross, K. N. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science* **2006**, *313*, 1929–1935.
- (764) Sirota, M.; Dudley, J. T.; Kim, J.; Chiang, A. P.; Morgan, A. A.; Sweet-Cordero, A.; Sage, J.; Butte, A. J. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data. *Sci. Transl. Med.* **2011**, *3*, 96ra77.
- (765) Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; Bork, P. Drug Target Identification Using Side-Effect Similarity. *Science* **2008**, *321*, 263–266.
- (766) Yang, L.; Agarwal, P. Systematic Drug Repositioning Based on Clinical Side-Effects. *PLoS One* **2011**, *6*, e28025.
- (767) Sanseau, P.; Agarwal, P.; Barnes, M. R.; Pastinen, T.; Richards, J. B.; Cardon, L. R.; Mooser, V. Use of Genome-Wide Association Studies for Drug Repositioning. *Nat. Biotechnol.* **2012**, *30*, 317–320.
- (768) Chiang, A. P.; Butte, A. J. Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses. *Clin. Pharmacol. Ther.* **2009**, *86*, 507–510.
- (769) Li, Y.; Agarwal, P. A Pathway-Based View of Human Diseases and Disease Relationships. *PLoS One* **2009**, *4*, e4346.
- (770) Suthram, S.; Dudley, J. T.; Chiang, A. P.; Chen, R.; Hastie, T. J.; Butte, A. J. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Comput. Biol.* **2010**, *6*, e1000662.
- (771) March-Vila, E.; Pinzi, L.; Sturm, N.; Tinivella, A.; Engkvist, O.; Chen, H.; Rastelli, G. On the Integration of in Silico Drug Design Methods for Drug Repurposing. *Front. Pharmacol.* **2017**, *8*, 298.
- (772) Byrne, R.; Schneider, G. In Silico Target Prediction for Small Molecules. In *Systems Chemical Biology*; Humana Press: New York, 2019.
- (773) Bellazzi, R.; Diomidous, M.; Sarkar, I. N.; Takabayashi, K.; Ziegler, A.; McCray, A. T. Data Analysis and Data Mining: Current Issues in Biomedical Informatics. *Methods Inf. Med.* **2011**, *50*, 536–544.
- (774) Margolis, R.; Derr, L.; Dunn, M.; Huerta, M.; Larkin, J.; Sheehan, J.; Guyer, M.; Green, E. D. The National Institutes of Health’s Big Data to Knowledge (BD2K) Initiative: Capitalizing on Biomedical Big Data. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 957–958.
- (775) Gottlieb, A.; Stein, G. Y.; Ruppin, E.; Sharan, R. PREDICT: A Method for Inferring Novel Drug Indications with Application to Personalized Medicine. *Mol. Syst. Biol.* **2011**, *7*, 496.
- (776) Zhang, P.; Agarwal, P.; Obradovic, Z. Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Blockeel, H., Kersting, K., Nijssen, S., Zelezny, F., Eds.; Lecture Notes in Computer Science; Springer, 2013, pp 579–594.
- (777) Xia, Z.; Wu, L.-Y.; Zhou, X.; Wong, S. T. Semi-Supervised Drug–Protein Interaction Prediction from Heterogeneous Biological Spaces. *BMC Syst. Biol.* **2010**, *4*, S6.
- (778) Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A Network Integration Approach for Drug–Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nat. Commun.* **2017**, *8*, 573.
- (779) Napolitano, F.; Zhao, Y.; Moreira, V. M.; Tagliaferri, R.; Kere, J.; D’Amato, M.; Greco, D. Drug Repositioning: A Machine-Learning Approach Through Data Integration. *J. Cheminformatics* **2013**, *5*, 30.
- (780) Oh, M.; Ahn, J.; Yoon, Y. A Network-Based Classification Model for Deriving Novel Drug–Disease Associations and Assessing Their Molecular Actions. *PLoS One* **2014**, *9*, e111668.
- (781) Liu, Z.; Guo, F.; Gu, J.; Wang, Y.; Li, Y.; Wang, D.; Lu, L.; Li, D.; He, F. Similarity-Based Prediction for Anatomical Therapeutic Chemical Classification of Drugs by Integrating Multiple Data Sources. *Bioinformatics* **2015**, *31*, 1788–1795.
- (782) Dai, W.; Liu, X.; Gao, Y.; Chen, L.; Song, J.; Chen, D.; Gao, K.; Jiang, Y.; Yang, Y.; Chen, J. Matrix Factorization-Based Prediction of Novel Drug Indications by Integrating Genomic Space. *Comput. Math. Methods Med.* **2015**, *2015*, 275045.
- (783) Luo, H.; Wang, J.; Li, M.; Luo, J.; Peng, X.; Wu, F.-X.; Pan, Y. Drug Repositioning Based on Comprehensive Similarity Measures and Bi-Random Walk Algorithm. *Bioinformatics* **2016**, *32*, 2664–2671.
- (784) Noeske, T.; Sasse, B. C.; Stark, H.; Parsons, C. G.; Weil, T.; Schneider, G. Predicting Compound Selectivity by Self-Organizing Maps: Cross-Activities of Metabotropic Glutamate Receptor Antagonists. *ChemMedChem* **2006**, *1*, 1066–1068.
- (785) Iorio, F.; Bosotti, R.; Scacheri, E.; Belcastro, V.; Mithbaokar, P.; Ferriero, R.; Murino, L.; Tagliaferri, R.; Brunetti-Pierri, N.; Isacchi, A.; et al. Discovery of Drug Mode of Action and Drug Repositioning from Transcriptional Responses. *P. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 14621–14626.
- (786) Carrella, D.; Napolitano, F.; Rispoli, R.; Miglietta, M.; Carissimo, A.; Cutillo, L.; Sirci, F.; Gregoretti, F.; Di Bernardo, D. Mantra 2.0: An Online Collaborative Resource for Drug Mode of Action and Repurposing by Network Analysis. *Bioinformatics* **2014**, *30*, 1787–1788.
- (787) Frey, B. J.; Dueck, D. Clustering by passing Messages Between Data Points. *Science* **2007**, *315*, 972–976.
- (788) Iorio, F.; Shrestha, R. L.; Levin, N.; Boilot, V.; Garnett, M. J.; Saez-Rodriguez, J.; Draviam, V. M. A Semi-Supervised Approach for Refining Transcriptional Signatures of Drug Response and Repositioning Predictions. *PLoS One* **2015**, *10*, e0139446.
- (789) Wu, Z.; Zhao, X.-M.; Chen, L. A Systems Biology Approach to Identify Effective Cocktail Drugs. *BMC Syst. Biol.* **2010**, *4*, S7.
- (790) Belkin, M.; Niyogi, P.; Sindhwani, V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *J. Mach. Learn. Res.* **2006**, *7*, 2399–2434.
- (791) Yan, X.-Y.; Zhang, S.-W.; Zhang, S.-Y. Prediction of Drug–Target Interaction by Label Propagation with Mutual Interaction Information Derived from Heterogeneous Network. *Mol. BioSyst.* **2016**, *12*, 520–531.
- (792) Chen, H.; Zhang, Z. A Semi-Supervised Method for Drug–Target Interaction Prediction with Consistency in Networks. *PLoS One* **2013**, *8*, e62975.
- (793) Mei, J.-P.; Kwok, C.-K.; Yang, P.; Li, X.-L.; Zheng, J. Drug–Target Interaction Prediction by Learning From Local Information and Neighbors. *Bioinformatics* **2013**, *29*, 238–245.
- (794) Dugger, S. A.; Platt, A.; Goldstein, D. B. Drug Development in the Era of Precision Medicine. *Nat. Rev. Drug Discovery* **2018**, *17*, 183–196.
- (795) Schadt, E. E.; Linderman, M. D.; Sorenson, J.; Lee, L.; Nolan, G. P. J. N. r. g. Computational Solutions to Large-Scale Data Management and Analysis. *Nat. Rev. Genet.* **2010**, *11*, 647–657.
- (796) Noor, A. M.; Holmberg, L.; Gillett, C.; Grigoriadis, A. Big Data: the Challenge for Small Research Groups in the Era of Cancer Genomics. *Br. J. Cancer* **2015**, *113*, 1405–1412.
- (797) Grossman, R. L.; Heath, A. P.; Ferretti, V.; Varmus, H. E.; Lowy, D. R.; Kibbe, W. A.; Staudt, L. M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112.
- (798) The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine Through an International Consortium. *Cancer Discov* **2017**, *7*, 818–831.
- (799) Tate, J. G.; Bamford, S.; Jubb, H. C.; Sondka, Z.; Beare, D. M.; Bindal, N.; Boutsidakis, H.; Cole, C. G.; Creatore, C.; Dawson, E.; et al. COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **2018**, *47*, D941–D947.

- (800) Jubb, H. C.; Saini, H. K.; Verdonk, M. L.; Forbes, S. A. COSMIC-3D Provides Structural Perspectives on Cancer Genetics for Drug Discovery. *Nat. Genet.* **2018**, *50*, 1200–1202.
- (801) Broniatowski, D. A.; Paul, M. J.; Dredze, M. National and Local Influenza Surveillance Through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. *PLoS One* **2013**, *8*, e83672.
- (802) pharmaphorum. <https://pharmaphorum.com/news/novartis-expands-virtual-clinical-trial-project-science-37/> (accessed June 20, 2019).
- (803) Chen, B.; Butte, A. Leveraging Big Data to Transform Target Selection and Drug Discovery. *Clin. Pharmacol. Ther.* **2016**, *99*, 285–297.
- (804) GNS Healthcare. <https://www.gnshealthcare.com/gns-healthcare-announces-collaboration-to-power-cancer-drug-development/> (accessed June 20, 2019).
- (805) EurekAlert! American Association for the Advancement of Science (AAAS). [https://www.eurekalert.org/pub\\_releases/2017-08/imiec081417.php](https://www.eurekalert.org/pub_releases/2017-08/imiec081417.php) (accessed June 20, 2019).
- (806) Zeng, T.; Zhang, W.; Yu, X.; Liu, X.; Li, M.; Chen, L. Big-Data-Based Edge Biomarkers: Study on Dynamical Drug Sensitivity and Resistance in Individuals. *Briefings Bioinf.* **2016**, *17*, 576–592.
- (807) Cision PR Newswire. <https://www.prnewswire.com/news-releases/berg-enters-into-an-agreement-with-sanofi-pasteur-to-identify-biomarkers-of-flu-vaccine-performance-300545009.html> (accessed June 20, 2019).
- (808) Lee, W.-S.; Ahn, S.-M.; Uhn, L.; Skjodt, N.; Zhavoronkov, A. Population Specific Biomarkers of Human Aging: A Big Data Study Using South Korean, Canadian, and Eastern European Patient Populations. *J. Gerontol A Biol. Sci. Med. Sci.* **2018**, *73*, 1482–1490.
- (809) Belsky, D. W.; Huffman, K. M.; Pieper, C. F.; Shalev, I.; Kraus, W. E. Change in the Rate of Biological Aging in Response to Caloric Restriction: CALERIE Biobank Analysis. *J. Gerontol., Ser. A* **2018**, *73*, 4–10.
- (810) Ellebrecht, C. T.; Bhoj, V. G.; Nace, A.; Choi, E. J.; Mao, X.; Cho, M. J.; Di Zenzo, G.; Lanzavecchia, A.; Seykora, J. T.; Cotsarelis, G.; et al. Reengineering Chimeric Antigen Receptor T Cells for Targeted Therapy of Autoimmune Disease. *Science* **2016**, *353*, 179–184.
- (811) Chatenoud, L. Precision Medicine for Autoimmune Disease. *Nat. Biotechnol.* **2016**, *34*, 930–932.
- (812) Stein, E. M.; DiNardo, C. D.; Polleyea, D. A.; Fathi, A. T.; Roboz, G. J.; Altman, J. K.; Stone, R. M.; DeAngelo, D. J.; Levine, R. L.; Flinn, I. W.; et al. Enasidenib in Mutant IDH2 Relapsed or Refractory Acute Myeloid Leukemia. *Blood* **2017**, *130*, 722–731.
- (813) Williams, T. J. Eotaxin-1 (CCL11). *Front. Immunol.* **2015**, *6*, 84.
- (814) Senn, H. M.; Thiel, W. QM/MM Studies of Enzymes. *Curr. Opin. Chem. Biol.* **2007**, *11*, 182–187.
- (815) Seabra, G. d. M.; Walker, R. C.; Elstner, M.; Case, D. A.; Roitberg, A. E. Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations Within the Amber Molecular Dynamics Package. *J. Phys. Chem. A* **2007**, *111*, 5655–5664.
- (816) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.
- (817) Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjan, U. N.; Duncan, P.; Rupp, M.; Müller, K. R.; Burke, K. Understanding Machine-Learned Density Functionals. *Int. J. Quantum Chem.* **2016**, *116*, 819–833.
- (818) Rupp, M. Machine Learning for Quantum Mechanics in a Nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.
- (819) Behler, J. Representing Potential Energy Surfaces by High-Dimensional Neural Network Potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.
- (820) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
- (821) Behler, J. Constructing High-Dimensional Neural Network Potentials: A Tutorial Review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.
- (822) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828–12840.
- (823) Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.
- (824) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.
- (825) Yao, K.; Parkhill, J. Kinetic Energy of Hydrocarbons as a Function of Electron Density and Convolutional Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 1139–1147.
- (826) Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. Orbital-Free Bond Breaking via Machine Learning. *J. Chem. Phys.* **2013**, *139*, 224104.
- (827) Mills, K.; Spanner, M.; Tamblyn, I. Deep Learning and the Schrödinger Equation. *Phys. Rev. A: At., Mol., Opt. Phys.* **2017**, *96*, 042113.
- (828) Liu, F.; Du, L.; Zhang, D.; Gao, J. Direct Learning Hidden Excited State Interaction Patterns from ab initio Dynamics and Its Implication as Alternative Molecular Mechanism Models. *Sci. Rep.* **2017**, *7*, 8737.
- (829) Häse, F.; Valleau, S.; Pyzer-Knapp, E.; Aspuru-Guzik, A. Machine Learning Exciton Dynamics. *Chem. Sci.* **2016**, *7*, 5139–5147.
- (830) Carpenter, B. K.; Ezra, G. S.; Farantos, S. C.; Kramer, Z. C.; Wiggins, S. Empirical Classification of Trajectory Data: An Opportunity for the Use of Machine Learning in Molecular Dynamics. *J. Phys. Chem. B* **2018**, *122*, 3230–3241.
- (831) Yao, K.; Herr, J. E.; Parkhill, J. The Many-Body Expansion Combined with Neural Networks. *J. Chem. Phys.* **2017**, *146*, 014106.
- (832) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with on-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
- (833) Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate Ab Initio Molecular Dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
- (834) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.
- (835) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B.; Ramprasad, R.; Gubernatis, J.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.
- (836) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
- (837) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated Materials Property Predictions and Design Using Motif-Based Fingerprints. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, 014106.
- (838) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3*, 2810.
- (839) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *93*, 115104.
- (840) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495–6502.
- (841) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; et al. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120–1127.

- (842) Ma, X.; Li, Z.; Achenie, L. E.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO<sub>2</sub> Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528–3533.
- (843) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.
- (844) Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 4934–4946.
- (845) Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery: Miniperspective. *J. Med. Chem.* **2014**, *57*, 7874–7887.
- (846) Zhang, W.; Pei, J.; Lai, L. Computational Multitarget Drug Design. *J. Chem. Inf. Model.* **2017**, *57*, 403–412.
- (847) Collins, K. D.; Glorius, F. A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nat. Chem.* **2013**, *5*, 597–601.
- (848) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725–732.
- (849) Corey, E. Computer-Assisted Analysis of Complex Synthetic Problems. *Q. Rev., Chem. Soc.* **1971**, *25*, 455–482.
- (850) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Nelson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921–1932.
- (851) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System—Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Model.* **1995**, *35*, 34–44.
- (852) Röse, P.; Gasteiger, J. EROS 6.0, a Knowledge Based System for Reaction Prediction—Application to the Regioselectivity of the Diels-Alder Reaction. In *Software Development in Chemistry* 4; Springer: Berlin, 1990.
- (853) Gelernter, H.; Sanders, A.; Larsen, D.; Agarwal, K.; Boivie, R.; Spritzer, G.; Searleman, J. Empirical Explorations of SYNCHEM. *Science* **1977**, *197*, 1041–1049.
- (854) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-Like Molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768–2777.
- (855) Fialkowski, M.; Bishop, K. J.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and Evolution of Organic Chemistry. *Angew. Chem., Int. Ed.* **2005**, *44*, 7263–7269.
- (856) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904–5937.
- (857) Peplow, M. Organic Synthesis: The Robo-Chemist. *Nature* **2014**, *S12*, 20–22.
- (858) Gothard, C. M.; Soh, S.; Gothard, N. A.; Kowalczyk, B.; Wei, Y.; Baytekin, B.; Grzybowski, B. A. Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry. *Angew. Chem., Int. Ed.* **2012**, *51*, 7922–7927.
- (859) Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. Parallel Optimization of Synthetic Pathways Within the Network of Organic Chemistry. *Angew. Chem., Int. Ed.* **2012**, *51*, 7928–7932.
- (860) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247–266.
- (861) Judson, P. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers*; Royal Society of Chemistry: Cambridge, 2009.
- (862) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-Aided Synthesis Design: 40 Years On. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2012**, *2*, 79–107.
- (863) Judson, P. *Knowledge-Based Expert Systems in Chemistry: Artificial Intelligence in Decision Making*; Royal Society of Chemistry: London, 2019.
- (864) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An Efficient Atom-Mapping Algorithm for Chemical Reactions. *J. Chem. Inf. Model.* **2013**, *53*, 2812–2819.
- (865) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56*, 2336–2346.
- (866) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **2017**, *3*, 434–443.
- (867) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. Doctoral Dissertation, University of Cambridge, 2012.
- (868) Lowe, D. *Chemical reactions from US patents (1976-Sep2016)*, [https://figshare.com/articles/Chemical\\_reactions\\_from\\_US\\_%20patents\\_1976-Sep2016/\\_/5104873](https://figshare.com/articles/Chemical_reactions_from_US_%20patents_1976-Sep2016/_/5104873) (accessed June 20, 2019).
- (869) Segler, M. H.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, *23*, 6118–6128.
- (870) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.
- (871) Segler, M. H.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966–5971.
- (872) Segler, M. H.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604–610.
- (873) Coulom, R. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. *International Conference on Computers and Games 2007*, *4630*, 72–83.
- (874) Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I.; Rohlfschagen, P.; Tavener, S.; Perez, D.; Samothrakis, S.; Colton, S. A Survey of Monte Carlo Tree Search Methods. *IEEE Trans. Comput. Intell. AI Games.* **2012**, *4*, 1–43.
- (875) Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research* **2018**, *62*, 729–754.
- (876) Niepert, M.; Ahmed, M.; Kutzkov, K. Learning Convolutional Neural Networks for Graphs. In *International Conference on Machine Learning*, 2016, 2014–2023.
- (877) Jain, A.; Zamir, A. R.; Savarese, S.; Saxena, A. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 5308–5317.
- (878) Mathea, M.; Klinspsohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inf.* **2016**, *35*, 160–180.
- (879) Klinspsohn, W.; Mathea, M.; Ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of Different Measures for Defining the Applicability Domain of Classification Models. *J. Cheminformatics.* **2017**, *9*, 44.