Enhancing preclinical drug discovery with artificial intelligence

R.S.K. Vijayan\textsuperscript{a}, Jan Kihlberg\textsuperscript{b}, Jason B. Cross\textsuperscript{a,*}, Vasanthanathan Poongavanam\textsuperscript{b,*}

\textsuperscript{a}Institute for Applied Cancer Science, MD Anderson Cancer Center, Houston, TX, USA
\textsuperscript{b}Department of Chemistry-BMC, Uppsala University, Uppsala, Sweden

Artificial intelligence (AI) is becoming an integral part of drug discovery. It has the potential to deliver across the drug discovery and development value chain, starting from target identification and reaching through clinical development. In this review, we provide an overview of current AI technologies and a glimpse of how AI is reimagining preclinical drug discovery by highlighting examples where AI has made a real impact. Considering the excitement and hyperbole surrounding AI in drug discovery, we aim to present a realistic view by discussing both opportunities and challenges in adopting AI in drug discovery.

Keywords: Artificial intelligence; Machine learning; Deep learning; Drug discovery

Introduction
Drug discovery is a long, complex, and high-risk process. It typically takes a staggering 10–15 years and costs up to US $2.8 billion on average, to develop a new drug, while an astonishing proportion (80–90%) of them fail in the clinic,\textsuperscript{1} with Phase II proof-of-concept (PoC) trials accounting for the most significant number of clinical failures. Although the number of new molecular entities (NMEs) approved by regulatory agencies, such as the US Food and Drug Administration (FDA), has increased over the past decade (2010–2019) compared with the prior decade, the cost of bringing a new drug to market has risen precipitously.\textsuperscript{1–3} The key drivers contributing to the increased cost of pharmaceutical innovation include investment lost from late-stage clinical attrition, an increasingly stringent regulatory system that sets a high bar for approval, and higher clinical trial costs, especially for

\* Corresponding authors. Cross, J.B. (jbcross@mdanderson.org), Poongavanam, V. (vasanthanathan.poongavanam@kemi.uu.se).

1359-6446/© 2021 The Author(s). Published by Elsevier Ltd.
https://doi.org/10.1016/j.drudis.2021.11.023This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

www.drugdiscoverytoday.com 967
pivotal trials. Given these realities, pharmaceutical and biotech companies are incentivized to innovate and adopt new technologies to improve productivity, cut costs, and ensure sustainability.

AI is shaping the evolution of entire industries, including health care (IBM Watson Health and Google’s DeepMind Health). Unsurprisingly, the biopharmaceutical industry also recognizes the potential value of AI and has shown keen interest in adopting AI-driven discovery platforms in the hope of streamlining R&D efforts, reducing discovery timelines and cost, and improving efficiency.\(^4\)–\(^5\) Major pharmaceutical companies have made significant investments in AI technology, including equity investments, acquisitions of, or partnerships with, AI-focused companies, building internal capabilities, or a combination of approaches. Partnerships appear to be focused on fast tracking the development of novel therapies, drugging ‘undruggable’ targets, broadening target portfolio through identification of novel targets, and improving the odds of clinical success.\(^6\) Big-tech companies, such as IBM, Microsoft, Amazon, and Google, which have competency and expertise in AI, are also making a foray into the drug discovery space.\(^7\) Public–private initiatives, such as the ATOM consortium (https://atomscience.org), have also been established with a mission of transforming drug discovery using data-driven modeling.

The AI technologies used today in drug discovery have evolved from earlier machine learning (ML) and cheminformatics concepts. For example, the application of ML to the development of quantitative structure–activity relationship (QSAR) models and expert systems for toxicity prediction has a longstanding history.\(^8\)–\(^9\) The widespread adoption of these techniques witnessed in recent times has been fueled by the advent of big data, advanced analytics, GPU-acceleration, cloud computing, algorithm development, and the democratization of AI toolkits.\(^3\)

The use of AI technologies is driving new opportunities across the drug discovery and development continuum, starting from target identification through to preclinical development (Fig. 1a). Evidence suggests that lack of clinical efficacy has been the foremost cause of attrition in clinical Phase II studies,\(^10\) highlighting that target selection remains one of the most crucial decision points in drug discovery. Consequently, there is a desire to improve the target selection process by applying AI techniques. AI-driven target discovery platforms can extract and synthesize target-relevant information from a large volume of complex, disparate multi-omics data, providing a better understanding of target biology, uncovering disease–target associations, thus identifying targets with a strong link to a disease. TargetDB is one such example that integrates publicly available data on a given target and uses an ML-based classification system to categorize target tractability.\(^11\) The approach and scoring system used within TargetDB provides useful criteria for ligandability assessment and prioritization of drug targets for development.

Once a target of interest has been identified and validated, the next stage in drug discovery is to identify high-quality chemical start points (hits) that bind to, and modulate the target. Although there are a range of hit-finding methods available, virtual screening (VS) is a cost-effective, and resource-sparing approach used to prioritize a subset of compounds for evaluation in a primary assay. The use of AI-driven approaches to improve the performance of VS is increasing.\(^12\) AI-powered VS campaigns have identified novel hits against seemingly difficult to drug targets,\(^13\)–\(^14\) thereby turning them into tractable drug targets.

To ensure that quality hits worthy of further consideration are progressed, computational methods have been used to identify, prioritize, and select hit compounds, a process referred to as hit triage.\(^15\) ML models are now being used to automate and improve the efficiency of this process.\(^16\)

Fast, accurate, and reliable prediction of binding free energies to enable VS and structure-based design remains a significant challenge, including rank-ordering of compounds from a VS. In recent years, ML-based scoring functions
trained on databases of protein–ligand complexes have shown great promise in improving hit rates during VS. Unlike traditional scoring functions, ML-based scoring functions can implicitly account for molecular interactions that are difficult to model, and are not constrained to any predefined functional form. With the advent of ‘make-on-demand’ libraries and screening collections breaking the billion compound barrier, conventional docking methods have become impractical. Active learning methods integrated with molecular docking offer an elegant solution for efficient exploration of chemical space through iterative screening.\(^{18-19}\)

The lead optimization (LO) phase is the most expensive and time-consuming phase in preclinical drug discovery.\(^{20}\) It is inherently a multiparameter optimization (MPO)\(^ {21}\) problem, with the goal of identifying compounds with an optimal balance of drug-like properties while maintaining sufficient potency. Hitting this ‘sweet spot’ is a challenge, because it involves simultaneous optimization of multiple and often competing objectives, such as safety, specificity, efficacy, and pharmacokinetics (PK) properties, while maintaining potency.\(^ {22}\)

LO involves iterative rounds of the design–make–test–analyze (DMTA) cycle (Fig. 1b), and reducing the number of iterations is crucial for accelerating the LO process. Generative chemistry that relies on AI-guided generative modeling for compound design has demonstrated success in reducing the number of iterations and designing compounds that meet the defined LO criteria.\(^ {23}\) Generative modeling platforms also integrate various predictive models for calculating various absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints to guide the design and selection of compounds with favorable properties that satisfy the defined LO criteria. In this way, generative chemistry can automate and shorten the ‘design’ phase of the DMTA cycle and offset individual cognitive biases during molecule ideation.

AI is also making headway in computer-aided synthesis planning (CASP), which is valuable in both hit identification and improving DMTA cycle efficiency.\(^ {24}\) AI-assisted synthesis planning helps chemists to objectively choose the most efficient and cost-effective synthetic route for a target molecule, thus accelerating the ‘make’ phase of the DMTA cycle. Automated continuous flow chemical synthesis is another emerging technology poised to revolutionize organic synthesis.\(^ {25}\) This technology opens new avenues by integrating smart automation and intelligent synthesis, thereby enabling fully autonomous synthesis.

Closing the loop of the DMTA cycle is the ‘analysis’ phase. To improve DMTA cycle efficiency, the data must be turned into knowledge to make better design suggestions for the next iteration. Given the sparse and non-uniform nature of the data encountered in drug discovery, the incorporation of sparse data AI methods, such as few-shot learning, for data analysis allows extracting valuable insights to inform the next round of the design cycle. Another practical application of AI is using deep imputation methods to handle the noisy, sparse, missing, and truncated data often generated in drug discovery.\(^ {26-28}\) Deep imputation methods combine DL and statistical imputation
methods to learn correlations between experimental endpoints and gain valuable information, even from minimal experimental data, to more accurately fill in missing experimental values.28 Such techniques can help establish assay–assay correlations and build multitarget QSAR models, which can be used for in silico off-target profiling against protein target families, such as kinases.

Translating preclinical discoveries into clinical practice in the form of new therapeutics is one of the biggest challenges in clinical development and, too often, clinical candidates are lost during translation. To bridge this ‘translational gap’, translational strategies are increasingly being integrated as early as LO to improve Phase II and Phase III clinical success rates, which is most evident in oncology drug discovery programs.29 To that end, translational biomarkers that demonstrate target modulation, target engagement, and confirm proof of mechanism (PoM) are used for de-risking clinical development. The ability of AI techniques to learn hidden and meaningful patterns by integrating large amounts of heterogeneous and high-dimensional omics data sets provides valuable insights for translational biomarker discovery.30 As innovations in AI technologies continue, the use of AI in drug discovery will also continue to grow.

AI toolbox for drug discovery

AI draws inspiration from diverse disciplines and brings together many technologies, such as ML, deep learning (DL), and data analytics. This has also led to a continually growing vocabulary used to describe AI. Although these terms are often muddled and used interchangeably, they have distinct meanings and relationships to one another, including properties such as data requirements, complexity, transparency, and capabilities. In general, AI is an umbrella term and considered a superset of ML, which itself is a superset of DL (Fig. 2a). Table 1 summarizes AI algorithms and their application in preclinical drug discovery.
ML algorithms are designed to build models that learn from problem-specific training data by identifying complex patterns and predicting outcomes on unseen data without being explicitly programmed. For these reasons, they are widely used in preclinical drug discovery and have been successfully used to predict bioactivity, ADMET related endpoints, and physicochemical properties with increased levels of accuracy. ML algorithms are broadly categorized as supervised, unsupervised, and reinforcement learning (RL) (Fig. 2b). The choice of the ML algorithm depends on many factors, including the data set and the type of problem.

Supervised learning methods use labeled data to train models and, once trained, the models can be used to predict outcomes on unseen data. These algorithms can handle both categorical and continuous data and are commonly used for classification and regression-based modeling approaches. Classification methods can be further categorized into binary, multiclass, and multilabel depending upon the specific classification tasks (Fig. 2c). Supervised learning methods have been shown to outperform inductive learning methods in multiple instances and have been used for similarity searching, predicting bioactivity, and other properties of interest. Kernel methods, such as support vector machines (SVMs), can map high-dimensional vector spaces, allowing for molecular featurization beyond classical molecular descriptors and fingerprints (e.g., MACCS keys and extended connectivity fingerprints). Supervised ML algorithms can also handle high-dimensional data to overcome the ‘curse of dimensionality’, and the collinearity problems often encountered in QSAR modeling, in addition to providing options for hyperparameter optimization when searching for the best model. A review of publications from leading pharmaceutical companies between 2014 and 2018 relating the use of AI in drug discovery revealed Random Forest (RF), SVMs, and other regression algorithms to be the most widely used ML techniques.

Conversely, unsupervised learning algorithms are trained on input data that are not labeled and are often used as a part of exploratory data analysis, such as clustering and dimensionality reduction. Some standard unsupervised learning algorithms include kappa-means clustering (k-mean), hierarchical clustering, principal component analysis (PCA), self-organizing map (SOM), and t-distributed stochastic neighbor embedding (t-SNE). In general, these methods are termed dimensionality

**TABLE 1**

| ML and DL algorithms and their application in preclinical drug discovery. |
|-----------------------------|-----------------------------|-----------------------------|
| **AI category** | **Algorithms** | **Task** | **Drug discovery applications** |
| **ML/supervised** | Decision tree | Classification | Binding affinity prediction; binding pose prediction; VS; PhysChem property prediction; ADMET |
| | SVM | Classification/regression | |
| | RF | Classification/regression | |
| | kNN | Classification | |
| | Naive Bayesian | Classification | |
| | Multiplier perception | Classification | |
| | GB | Classification | |
| | ANN | Classification/regression | |
| | Linear regression (e.g., MLR) | Regression | |
| | Polynomial regression | Regression | |
| | Partial least squares regression | Regression | |
| | Support vector regression | Regression | |
| **ML/unsupervised** | PCA | Clustering/dimensionality reduction | VS; data visualization; dimensionality reduction; molecular pattern recognition |
| | k-mean clustering | Clustering | |
| | SOM | Clustering | |
| | Hierarchical clustering | Clustering | |
| | t-SNE | Clustering | |
| | Singular value decomposition | Dimensionality reduction | |
| | Linear discriminant analysis | Dimensionality reduction | |
| | Multidimensional scaling | Dimensionality reduction | |
| | Partial least squares regression | Dimensionality reduction | |
| | Principal component regression | Dimensionality reduction | |
| | Gaussian mixture models | Probability distribution | |
| | Association rule learning (e.g., FP-Growth) | Pattern recognition | |
| **ML/RL** | Q-Learning | Generative modeling | De novo molecular design; VS; docking; MPO |
| | GENTRL | | |
| | ReLeASE | | |
| **DL** | CNNs | Representation-learning techniques | Protein structure prediction; binding affinity; binding pose prediction; de novo molecular design; focused library generation; PhysChem property prediction; ADMET; MPO; synthesis planning |
| | RNNs | | |
| | Deep belief networks (DBN) | | |
| | Stacked autoencoders | | |
| | LSTMs | | |
| | Deep Boltzmann machine (DBM) | | |

*a* Binary classifier.

*b* Multiclass classifier.

*c* Multilabel classifier.

**KEYNOTE (GREEN)**

Drug Discovery Today • Volume 27, Number 4 • April 2022

www.drugdiscoverytoday.com 971
reduction techniques and provide a means for projecting high-dimensional data into a low-dimensional space for visualization. They are often used in drug design for developing QSAR models (PCA-MLR modeling), designing screening libraries, clustering, data exploration, and visualizing the chemical space of large compound libraries. Several supervised learning algorithms, such as SVM and neural networks (NNs), can also support unsupervised learning.34

Unlike supervised and unsupervised learning, RL systems continuously interact with the environment using feedback from previous actions and experiences to achieve their goals. Each time a RL agent performs an action, it utilizes an objective function that is rewarded if the output is acceptable, and penalized when it is not. The goal of a RL algorithm is to identify the optimal policy to maximize the reward function. RL algorithms, such as Generative Tensorial Reinforcement Learning (GENTRL),37 and Reinforcement Learning for Structural Evolution (ReLeaSE), have been used for designing molecules with desired properties during generative modeling.38

DL is a subset of ML and belongs to a broader family of artificial neural network (ANN) algorithms. It is currently the state-of-art AI technology and can be described as a class of representation-learning techniques. ANN algorithms are loosely inspired by the structure of the human brain. Accordingly, ANN architecture contains many processing elements, called neurons, that are organized into multiple layers. The network comprises input nodes and a layer of output nodes connected by layer of hidden nodes (Fig. 2d). Each hidden node has an associated weight, activation function, and bias function that transforms the input data to make predictions. The adjective ‘deep’ in DL refers to an ANN with many layers, and the number of hidden layers signifies the depth of the network. DL methods contains several hidden layers as opposed to traditional ‘shallow-learning’ ML methods, which usually contain one or two hidden layers. Whereas DL methods use deep and specialized architectures to learn and extract higher-level features in an automated fashion from unstructured data, they also require a huge amount of training data. Another key difference between DL and shallow-learning ML algorithms is that DL algorithms scale with the data, whereas shallow-learning ML algorithms converge at a certain level of performance. Some popular DL architectures used in drug discovery include convolutional neural networks (CNNs), graph CNNs (GCNNs), autoencoders (AE), and recurrent NNs (RNNs).

CNNs are the most-utilized DL method in drug discovery.7 A CNN architecture comprises of multiple layers of neurons, each of which is fully connected to all neurons in the preceding layer. It usually contains several convolution layers and pooling layers occurring in an alternating fashion that are capable of learning any highly nonlinear function. Deep CNN models trained on 3D atomic grids extracted from experimental protein–ligand complex structures are now being used for structure-based VS and property prediction.34–35 They have been shown to successfully model the complex, nonlinear phenomenon of a small molecule binding to the protein,35 and have demonstrated significant improvement for property prediction.40 Other DL architectures, such as RNNs, have also been used for generative modeling and focused library generation.41

Successful applications of AI in drug discovery
Over the past few years, several reviews have been published that highlight the emerging role of AI in drug discovery.5,42–45 Hence, this review would focus on highlighting a few exemplary studies where AI has made a real impact in small-molecule drug discovery.

Structural enablement of drug targets and binding site comparisons
The availability of atomic-resolution structural information of small molecules binding to drug targets provides opportunities for structure-guided hit identification (structure-based VS), fragment screening (fragment-based drug discovery; FBDD), and ligand optimization (structure-based drug design; SBDD). Structural information of the target also provides insights into selectivity drivers, resistance mechanisms, mode of action, allosteric pocket identification, and ligandability assessment for novel drug targets.44–45 Despite technological advancements in X-ray crystallography, NMR spectroscopy, and single-particle cryo-EM, there is structural coverage for only 35% of the human proteome. In many instances, this structural coverage is often limited to a single structural domain of a multidomain protein. Hence, there is still a gap between the number of known protein sequences and the number of experimentally solved structures.46 Importantly, structural coverage for pharmacologically relevant protein target families, such as G-protein-coupled receptors (GPCRs) and ion channels, remains underrepresented in databases, such as the Protein Data Bank (PDB).47

An alternative approach to protein 3D structure generation in the absence of experimental structure is the use of computational structure prediction methods. Homology modeling has been the traditional approach to bridge the sequence–structure gap. It predicts the 3D structure of an unknown (target) protein based on the experimental structure of a homologous (template) protein given its amino acid sequence. It has been shown that homology-modeled structures with sequence identity down to 30% are generally suitable for SBDD.48 For proteins that lack a homologous structure, accurate structure prediction remains a challenge; however, advances in DL-based methods and the integration of coevolution data into modeling have invigorated the field of protein structure prediction.49 DL-based algorithms, such as CNN, RNN, variational autoencoders (VAEs), and generative adversarial networks, have demonstrated improved success in protein structure prediction even in the absence of a template structure.50

The use of DL methods for protein structure prediction took center stage with the remarkable success of a deep convolutional residual network (ResNet)-based program called AlphaFold2 at the CASP14 competition.51 AlphaFold2, developed by DeepMind technologies,52 used a DNN architecture trained on 170 000 protein structures from the PDB to predict the distribution of distances between pairs of amino acids and torsion angles between chemical bonds that connect those amino acids in a protein. In addition, it uses evolutionary information derived from multiple sequence alignment and an end-to-end folding method for structure prediction. The methodology and the architecture behind AlphaFold2 was recently published.53
To gauge structure prediction accuracy, CASP uses a Global Distance Test (GDT) metric that quantifies the residue correspondence between the model and the experimental structure. A score of 90 GDT implies the prediction accuracy to be comparable with experimental methods (Fig. 3); AlphaFold2 achieved a median score of 92.4 GDT for all targets. Thus, the results from the CASP14 competition revealed that DL methods could achieve impressive levels of accuracy comparable to experimental structures. DeepMind, in partnership with EMBL-EBI, has made freely available to the scientific community the 3D structures predicted by AlphaFold2, which offer structural coverage for 98.5% of the human proteome. Inspired by the idea and the success of AlphaFold2, an academic team led by David Baker also developed a three-track NN program dubbed RoseTTAFold.

Although these developments signify advancements in protein structure prediction, it is too early to state that AI has cracked the protein-folding problem or its impact on drug discovery will be transformative. AlphaFold2 was trained on over 170,000 protein structures from PDB, and any learning model is only as good as the data it was trained on. Current estimates of the number of folds in the PDB based on the SCOP version 2 database is 1388, whereas the number of folds in nature is predicted to somewhere between 4000–10,000. Hence, there are many novel folds, topologies, and architectures unseen in PDB, and there is also considerable redundancy in both sequence and protein families in this database. Furthermore, predicting the structure of multidomain proteins, multimeric protein complexes, and membrane proteins might be a harder problem to solve using AI. Nevertheless, DeepMind’s technology promises to advance structural biology and de novo protein design, and to drive drug discovery.

Comparing protein-binding pockets at the structural genomics scale is a valuable exercise in structure-based drug design. It provides information that can help understand selectivity, predict off-target liabilities, provide insights into drug repurposing, and aid protein function annotation. Traditional pocket comparison methods use representations, such as graph theory, geometric hashing, typed triangles, spherical harmonics, and physicochemical properties of the binding site atoms, to compute the sequence-free similarity between binding sites. These intuition-based featurization schemes can introduce human biases and are generally not scalable across thousands of binding sites.

The introduction of DNN algorithms has enabled the building of powerful voxel-based feature representations that can encode molecular properties and vectorize the binding sites into descriptor vectors. One such implementation is DeeplyTough, which uses 3D steerable CNNs for comparing binding sites in an alignment-free manner. This is accomplished by encoding 3D representations of protein pockets into descriptor vectors, which can be used for computing pairwise Euclidean distances to quantifying pocket similarity. It was trained on the TOUGH-M1 data set, which is a nonredundant and representative data set of small-molecule binding pockets with approximately 1 million data points. It includes a positive subset comprising different proteins that bind chemically similar ligands, and a negative subset, containing different proteins that bind chemically dissimilar ligands. Performance was evaluated against two independently constructed datasets (Vertex and ProSPECTs), with DeeplyTough demonstrating results competitive with existing approaches but with a reduced runtime.
Augmenting virtual screening using AI

VS is a computational technique that offers a complementary and cost-effective approach for hit identification relative to experimental screening methods, such as high-throughput screening. Instead of physically screening every compound from the screening collection, VS uses computational techniques to prioritize a subset of compounds for evaluation in a primary assay.

The increasing size of ‘make-on-demand’ screening libraries, and the expanding number of high-value, challenging drug targets identified from functional genomics screening present a significant challenge for conventional VS techniques. Hence, AI methods that augment VS approaches and help efficiently explore the chemical space for hit identification have garnered considerable attention in drug discovery.

Ligand-based virtual screening

Ligand-based VS (LBVS) techniques aim to identify active compounds from a chemical library based on the principle of molecular similarity. They include similarity searching, pharmacophore screening, shape matching, and predictive modeling.

Predictive modeling for VS is an extension of the classical QSAR modeling paradigm. Classical QSAR uses statistical data-modeling methods on a congeneric series to build explanatory models that quantify SAR trends in a retrospective manner. Access to a large volume of chemogenomics data (PubChem’s BioAssay and the ChEMBL database) and advances in ML and DL algorithms that can handle large data sets have provided new opportunities for QSAR modeling as a VS technique. Consequently, many successful applications of QSAR-based VS workflows for hit identification have been reported. Zhang et al. described the successful implementation of an ML-based QSAR workflow for VS that led to the discovery of novel antimalarial agents. The authors used two ML algorithms (SVM and kNN) to develop a binary classifier model (active or inactive), which was trained using 3133 compounds with known antimalarial activities. The QSAR models were used to carry out VS against the ChemBridge database and resulted in the selection of 174 compounds for a follow-up screening in *Plasmodium falciparum* growth inhibition and cellular assays. Experimental validation revealed 25 of the selected compounds to be active, yielding a hit rate of 14.2%, with the most potent hit having an EC50 value of 95.6 nM. Subsequently, many studies have reported the application of ML and DL-based QSAR workflows as promising VS tools.

Over the past decade, there has been a shift to web-based cheminformatics workbenches that streamline and automate ML- and DL-based QSAR workflows for VS. Liu et al. developed a user-friendly open-source web server called DeepScreening, which allows users to build and validate RNN models using either ChEMBL bioactivity data or user-provided data sets for VS. DeepScreening also provides prebuilt DNN models for 1251 targets based on the bioactivity data curated from ChEMBL 24. This user-friendly interface and the availability of prebuilt QSAR models allow QSAR experts and nonexperts to perform VS against a specific target of interest. DpubChem is another open-source web server that uses ML approaches to derive categorical QSAR models using PubChem data.

Whereas chemogenomic databases such as PubChem and ChEMBL provide sufficient bioactivity data to build models, there are still significant pitfalls associated with using these resources. Primary among these issues is the presence of bioactivity data from heterogeneous sources and an imbalanced ratio of active to inactive compounds for a given target. This makes the generalized use of QSAR-based workflows for VS more difficult to implement using public data sets compared with other virtual screening methods.

Structure-based virtual screening

A common computational strategy applied in SBDD is molecular docking. It plays a crucial role in driving many structurally enabled drug discovery programs, spanning from hit identification to LO and binding mode prediction. The docking process involves predicting the bound ligand conformation (pose prediction) within a binding site, followed by an estimation of its binding affinity (scoring).

Given significant progress in DL, the application of AI methods is becoming more commonplace in SBDD. Unlike shallow-learning AI methods that rely on feature engineering, DL can automatically learn and extract features from 3D structural data. Thus, DL methods, which are popular in image recognition, are now being applied to extract and generalize structural features from protein–ligand complexes through multilayer feature extraction. This opens opportunities for using AI methods in SBVS, binding mode prediction, and binding affinity estimation. The use of ML and DL algorithms for pose prediction and scoring during docking has been demonstrated to have superior performance in terms of scoring power (ability to rank order compounds based on binding affinities), docking power (ability to distinguish native poses from decoy poses), and screening power (ability to distinguish binders from nonbinders).

AtomNet was the first structure-based application that used a deep CNN framework to predict binding affinity. It uses a 3D grid approach to encode the environment of each atom in the binding site into voxelized feature vectors and was trained on a ChEMBL data set containing 78 000 actives and 2 000 000 decoys, spanning 290 targets. Interestingly, benchmark studies carried out using the DUD-E benchmark data set showed impressive performance, with AtomNet achieving an area under the curve (AUC) greater than 0.9 on 57.8% of the targets, far surpassing the conventional docking methods.

Accurate binding mode prediction of the ligand using docking is a key challenge in SBDD. Although binding free energy is a macroscopic observable that involves a ratio of partition functions between two states (bound and unbound), it is commonly assumed that the binding mode observed using experimental techniques, such as X-ray crystallography, corresponds to the lowest energy pose. Hence, most classical scoring functions that are parameterized to recapitulate binding affinities are in turn used to select the top-scoring docked pose as its predicted binding mode. However, accurate binding energy estimation using scoring functions remains a significant challenge, and this often results in ‘hard failures’, implying that the predicted binding modes do not correspond to native or near-native binding modes.

The first attempt to develop a task-specific scoring function for binding pose prediction using ML was carried out by Ash-
tawy and Mahapatra. They showed that various ML algorithms that can map structural and physicochemical information from protein–ligand complexes could distinguish native and near-native docked poses from decoy poses. The best task-specific ML scoring function displayed improved docking power (>14%) over classical scoring functions. A similar implementation that uses a 3D-CNN, called DeepBSP, can predict the root mean square deviation (RMSD) of a predicted pose relative to its native binding pose. The authors trained their model against a data set that contained 11,925 native complexes and more than 165,000 Autodock Vina docked decoy poses, and benchmarked their model using the CASF-2016 benchmark data set. Their findings revealed that scoring Autodock Vina generated poses with DeepBSP displayed an improved docking power relative to the hybrid knowledge and empirical-based scoring function available within the application.

In a prospective context, Adeshina et al. used ‘vScreenML’ for VS that led to the discovery of an AChE inhibitor with an IC50 value of 280 nM (Ki = 173 nM). vScreenML is built on an XGBoost framework and uses a classifier approach to categorize docked poses as actives or decoys. The authors attributed its performance to the unique nature of the training data set, D-COID, which includes both native complexes that are representative of drug-like compounds and decoy complexes generated by molecular modeling. The inclusion of decoy complexes in the training data significantly improved the classification accuracy of the vScreenML scoring function.

Recently, a new class of scoring functions, developed based on ML and DL algorithms, has been gaining popularity. Many common ML and DL architectures have been used for the development of ML-based scoring functions, including SVM, RF, kNN, gradient-boosting decision tree (GBDT), and 3D deep-CNN. Representative examples include Pafnucy, Onion-Net, RFscore v3, NNscore2.0, BgN(BsN)-Score, and ΔvinaRF. Onion-Net and Pafnucy use a 3D CNN approach for featurization of protein–ligand complexes. ML-based scoring functions have been shown to outperform classical scoring functions in various benchmark studies. Unlike traditional scoring functions, ML-based scoring functions are nonparametric because they do not have a predetermined functional form that approximates the underlying physics of molecular recognition. Instead, they are trained from experimental data sets that contain both protein–ligand structural data and binding affinity data, such as PDBbind and Binding MOAD. Hence, they implicitly account for molecular interactions that are had to model explicitly. ML-based scoring functions can be used either to rescore docked poses generated by an external docking program or integrated within a docking program to help in guiding the sampling of poses. A detailed discussion of ML-based scoring functions is beyond the scope of this review, and the reader is referred to several excellent reviews on this topic.

In addition, there are several studies that report the use of ML and DL-based scoring functions that lead to the identification of experimentally validated hits during VS campaigns. As an example, an ML-based scoring function, called MIEC-SVM, which combines molecular interaction energy components (MIEC) and SVM, was used to screen the Specs vendor database and led to the identification of a novel class of ALK kinase inhibitor. The use of MIEC-SVM to rescore Autodock 4.2 poses

![Diagram](https://example.com/diagram.png)
yielded a hit rate of 14% compared with a hit rate of 6% using the native scoring used by Autodock.

Although ML-based scoring functions offer an improvement over classical scoring functions, the interpretability of ML-based scoring functions is not straightforward and one needs to be cautious of issues during benchmarking and validation of ML-based scoring functions. Flaws in the design of benchmark data sets and improper splitting of the data set into training and test data can yield overly optimistic performance estimates. Data leakage happens if the benchmark data set happens to have the information present in the training data leading to inflated performance during validation. Overfitting of ML occurs when models show high accuracy in the training set but fail to generalize on unseen data set. Gabel et al. reported on two ML-based scoring functions, RF-IChem and SVM-Ichem, which were used to predict pKi values for 195 diverse protein–ligand X-ray structures. The ML-based scoring functions were found to be insensitive to the position, orientation, and conformation of the docked pose, suggesting that they had overestimated accuracies from overfitting. Conversely, the Surflex-Dock scoring function behaved as expected, being logically sensitive to changes in pose conformation. This reminds us of the downside of developing computational tools with ‘black box’ characteristics and the potential to introduce artifacts due to overfitting.

**Active learning docking**

As the size of make-on-demand libraries continues to grow, so does the need for computational tools that can efficiently navigate the chemical space during VS. Although the debate whether ‘bigger is better’ or ‘smaller is better’ in VS continues, docking of ultralarge libraries for hit identification is gaining popularity. The colossal size of such libraries poses a challenge for docking programs, which can rarely perform brute force docking exceeding 100 million molecules.

Integrating an active learning algorithm with molecular docking offers an elegant solution for scaling up the screening of ultralarge libraries. Active learning docking, in general, begins by docking a small subset of the entire library, then uses the results to train an ML model to predict docking scores for the rest of the compounds in the library. Top-scoring compounds from the ML model are then docked, and the ML model is updated with the new data. The whole process is repeated iteratively until the ML model converges. Thus, active learning docking, while preserving the fidelity of brute force docking, helps identify the highest-scoring compounds from ultralarge compound libraries at a fraction of the time required by conventional docking. An overview of this process is shown in Fig. 4.

One such implementation of active learning docking is ‘Active Learning Glide’, which integrates the docking program Glide and the ML algorithms available from the open-source framework DeepChem. Graff et al. demonstrated that a molecular-pool based active learning-guided docking approach was able to retrieve most of the top-scoring compounds in a virtual library at a fraction of the computational cost of brute force docking.

**Generative chemistry**

The concept of using computational methods for compound ideation has a rich history. Early structure-based de novo design approaches involved automated and incremental construction of ligands within the receptor binding site. Programs such as LUDI identify potential interaction sites in the binding pocket and build molecules from a predefined set of organic fragments that sterically and electronically complement the protein binding pocket. Inverse-QSAR modeling is another de novo molecular design approach that seeks to design molecules with a specific activity or property by inversely mapping the molecular descriptor from a preconstructed quantitative structure–activity/property relationship (QSAR/QSPR) models. In general, solving the inverse-QSAR problem is complex, because reconstructing molecular structures based on molecular descriptor information provided by a forward QSAR/QSPR model is challenging. A practical problem that precluded the widespread adoption of these classical de novo design approaches is the limited synthesizability and poor drug-like properties of the designed molecules.

Over the past few years, the use of AI-based generative modeling algorithms for de novo molecular design has gained in popularity because they can overcome issues encountered with classical de novo design approaches. Generative chemistry relies on the use of modern AI-based generative modeling tools to generate synthetically tractable compounds with drug-like properties while satisfying the desired target property profile. Based on a data-driven approach, generative modeling algorithms learn the underlying nonlinear distribution between molecular structures, their biological activity, and physicochemical properties from a large volume of data to inform compound design. AI-guided generative modeling platforms, in short, perform compound ideation, prediction, and selection of compounds with favorable properties. Several DL architectures, such as VAEs, generative adversarial networks (GANs), RL, and RNNs, have been applied to de novo molecular design. Current generative modeling methodologies can also be categorized depending on the underlying method used for molecular featurization. Whereas most first-generation generative modeling methods used fingerprints and SMILES strings to encode molecular structures, newer approaches, such as molecular graphs and fragment-based methods, are becoming increasingly popular.

One such study to establish proof-of-principle for implementing deep GANs in generative modeling was reported by Kadurin et al. An example demonstrating the utility of RNNs in generative modeling was also reported by Segler et al. The first report of the successful application of RNN containing long short-term memory (LSTM) cells for de novo molecular design was reported by Gupta et al. The generative LSTM model was trained against the ChEMBL22 database to generate novel molecules that could modulate retinoid X receptors (RXRs) and peroxisome proliferator-activated receptors (PPARs). The fine-tuning process involved training against a data set containing 25 RXR and PPAR modulators (agonists and partial agonists). The on-target activity of the generated compounds was predicted using an ML model, with four of the five top-ranked compounds selected for synthesis showing activity in a cell-based assay (Fig. 5a, 1–4). Two were found to be PPAR agonists, and two were dual inhibitors of PPAR and RXR, displaying EC50 values ranging from low to double-digit mM. Although the compounds were not extensively characterized, these findings demonstrate the
ability of generative AI to deliver synthetically tractable, novel bioactive molecules that satisfy design objectives.

Another generative modeling study that attracted significant press coverage at the time was the use of a GENTRL model by Zhavoronkov et al., which led to the discovery of potent kinase discoidin domain receptor (DDR1) inhibitors in only 21 days.37 The molecule (6) designed by the GENTRL approach is shown in Fig. 5b and is compared with the parent molecule (5) and other DDR1 inhibitor (7). The authors trained their generative model in a semisupervised fashion using an objective function that rewards synthetic feasibility, on-target activity, and novelty. Of the 30,000 molecules proposed by the generative model, six were subsequently synthesized and tested. Four compounds were found to be active in biochemical assays, and two were active in a cell-based assay, with the best compound (6) having an IC₅₀ value close to 10 nM in both biochemical and cell-based assays. Although this study demonstrated the ability of generative modeling to identify a nanomolar hit compound, concerns have been raised about the novelty of the molecules.112 The best molecule designed was strikingly similar to the marketed multityrosine kinase inhibitor ponatinib. In addition, selectivity profiling of the compound against the broader kinome was not established, which called into question the clinical value of the compound. The authors responded to these criticisms, stating that the study was meant to demonstrate the potential of generative modeling technology and not intended at identifying a clinical candidate compound.113

Recently, Perron et al. gave an account of what might be the first report of a successful application of generative modeling in solving a MPO problem.114 An LSTM generative model coupled to a RL method, on an undisclosed target, was used to design 150 compounds that were predicted to meet all the defined LO criteria. The training data set included 881 molecules with 11 sets of associated assay data, including on-target activity, off-target activity, and ADMET endpoints. None of the compounds in the training data satisfied all the property and potency criteria.
Twenty compounds generated by the model were shortlisted for synthesis based on different criteria; nine of these compounds failed during synthesis, the remaining 11 were synthesized and profiled, and one of these satisfied all the eleven LO criteria (Fig. 5c).

Although generative chemistry is becoming increasingly popular, emphasis should also be placed on the rigorous validation of generative models. Assessment methods for generative modeling should include the application of distribution learning benchmarks, synthetic validity, novelty, compound quality, goal-directed objectives, as a part of their evaluation framework. Open-source standardized benchmarking platforms, such as Molecular Sets (MOSES) and GuccaMol, could serve as a framework for benchmarking generative modeling methods.

**In silico ADMET prediction**

The observation that poor pharmacokinetics of drug candidates were an important cause of clinical attrition in the late 1990s brought about a paradigm shift within the pharmaceutical industry. It witnessed the emergence of several property-based drug-likeness rules, such as Lipinski’s Ro5, and many developability metrics to control compound properties during LO. In addition, the establishment of miniaturized, high-throughput in vitro ADMET profiling assays resulted in the parallel evaluation of efficacy and ADMET during earlier stages of drug discovery. In silico ADMET modeling is intended to assist project teams in the design and selection of novel compounds with superior ADMET properties and in directing experimental resources to the most favorable compounds, thereby reducing the overall number of compounds that need to be synthesized and profiled. Over the years, pharmaceutical companies have deployed many global in silico ADMET models in their discovery pipelines. A representative list of such in silico ADMET models is provided in Fig. 6a.

Early work in ADMET modeling used linear regression methods, such as those used by Hansch and Free-Wilson analysis. However, with the development of ML algorithms and the availability of large-scale homogenous ADMET data, in silico ADMET modeling transitioned toward ML-based predictive models developed using Bayesian neural networks, RFs, and SVMs. These ML algorithms are suitable for predicting endpoints that have a complex and nonlinear relationship.

The use of DNN methods for in silico modeling of ADMET endpoints gained popularity following the Kaggle ‘Merck Molecular Activity Challenge’ conducted in 2012. The Kaggle competition was meant to examine how well ML methods can predict 18 different ADMET endpoints using data sets of various sizes (2000–50 000 molecules) derived from Merck’s internal data. The winning entry used an ensemble approach that included DNN, gradient-boosting machine (GBM), and Gaussian process (GP) regression methods. Merck researchers released a follow-up study comparing the performance of DNN models to that of RF models, and demonstrated that DNN models outperformed RF in most cases. Similarly, in the Tox21 data challenge, conducted by the NIH to compare computational models for toxicity prediction, DL models excelled and outperformed shallow-learning ML models.

A unique characteristic of DNNs is their ability to simultaneously train NNs that combine different endpoints within a single model. Using a learning technique called inductive transfer...
learning, multitask DNN trains data sets corresponding to different ADMET endpoints and combines them into a single model. The rationale behind multitask DNNs is to enable faster learning and improved model accuracy by sharing their representation internally.129 Most multitask DNN models used for modeling ADMET endpoints use a ‘Hard’ parameter sharing approach, which implies sharing of the hidden layer between all the tasks.130 A general data-driven ADMET model building process is illustrated in Fig. 6b.

In an accuracy performance benchmark study carried out against 31 assay data sets, Evan et al. showed that multitask DNNs were more accurate in predicting ADMET endpoints over single-task DNNs and shallow-learning ML methods, such as RF.131 Scientists from Sanofi–Aventis AG also reported the successful implementation of predictive multitask DNN model into their in silico ADMET workflows.132 By applying an alternate multitask learning method to transfer features between data sets, they developed multitask models for predicting metabolic liability/clearance, Caco-2 permeability, and logD<sub>7.4</sub>. They also reported that multitask DNN methods, although outperforming single-task DNN methods in many instances, showed poorer performance compared with single-task DNNs for certain endpoints. Hence, combining mechanistically unrelated endpoints in a multitask model could lead to poor performance, since the information shared between the tasks might not be correlated.133 Hence, a priori assumption of the predictive advantage of multitask DNN over single-task DNNs is a challenge and both approaches need to be evaluated when developing predictive ADMET models.

Although predictive modeling has a significant role in selecting compounds with superior ADMET properties, they are often insufficient for guiding compound ideation during the lead optimization phase. Extracting tacit knowledge from the corpus of information generated from prior discovery programs can be used as a source of idea for compound design. One such widely used concept in medicinal chemistry is termed Matched Molecular Pair (MMP) analysis.134 MMP is defined as a pair of molecules that differ only by a well-defined structural transformation that is associated with a relative change in a property value.135 Traditionally, MMP analysis was carried out by analyzing the frequency of one-to-one structural transformations. With advances in AI technology and molecule fragmentation algorithms, MMP analysis of large data sets in an automated fashion is now feasible. A notable implementation is MCPairs,136 which uses an unsupervised ML approach to mine in vitro ADMET data integrated from three different pharmaceutical companies (AstraZeneca, Genentech, and Roche). The use of AI and the availability of large-scale data help develop next-generation MMP platforms that offer practical solutions to address ADMET issues using explainable AI.

**Computer-aided synthesis planning**

The use of computer-aided synthetic planning (CASP) dates back to the pioneering work of E.J. Corey,137 who formalized the concept of ‘retrosynthetic analysis’ during the late 1960s.138 Retrosynthetic analysis refers to a technique that involves the deconstruction of a target molecule into its simple, readily available starting materials by sequential disconnections and functional group interconversions. CASP programs incorporate the idea of retrosynthetic analysis and help synthetic organic chemists select the most efficient and cost-effective synthetic route.24,139 They can also be used for selectivity and side product prediction and for recommending and evaluating reaction conditions. The use of AI has revitalized the field of computer-aided synthesis planning, and technological developments over the years have been well reviewed in recent literature.24,140 Therefore, we limit the scope of our review to the application of CASP in the context of drug discovery. AI-assisted synthesis planning tools help chemists augment their synthetic chemistry knowledge by recommending viable synthetic routes. They also help chemists make better decisions, thereby improving efficiency and productivity by reducing synthesis failures.141 Ultimately, this accelerates the ‘make’ phase of the DMTA cycle in drug discovery.

Computer-aided synthetic route-planning strategies generally fall into two broad categories: rule or template-based methods and template-free methods. Rule-based methods use expert coded rules and heuristics extracted from reaction databases and literature for suggesting synthetic routes. In rule-based methods, the reaction rules are extracted and codified manually. One such example of retrosynthetic software, which uses a library of expert encoded rules for chemical synthesis planning, is Synthia (formerly Chematica). A limitation with such rule-based methods is its inability to scale with the exponential growth of chemical literature, and its knowledge base is limited because complete coverage is unlikely. To address these limitations, automated rule-based methods were developed for extracting reaction rules from reaction data sets using computational approaches. Such automated, rule-based methods use template extraction algorithms that rely on atom-mapped reaction examples in the form of SMIRKS patterns for extracting transformations from reaction data sets. Two significant limitations include the high computational cost involved with the calculation of subgraph isomorphism and the lack of chemical intelligence.142–143 An alternate rule-based approach developed in recent years for extracting reaction rules is the application of data-driven DL techniques. A notable example is a pioneering study by Segler et al.109 who used a neural-symbolic approach to extract retrosynthetic rules from the Reaxys database autonomously without expert input. These rules were then subjected to reaction prediction in combination with a modern Monte-Carlo tree search algorithm to select the most promising retrosynthetic steps.

An orthogonal approach to the classical rule-based methods is the use of template-free methods for reaction prediction and retrosynthetic transformations. Template-free methods draw inspiration from Natural Language Processing (NLP) and treat forward or retrosynthetic prediction as a neural machine translation problem.144 Given that molecules can be represented as SMILES strings, each chemical reaction can be encoded as sentences and treated as a chemical linguistics translational problem. Liu and coworkers proposed the first template-free model for retrosynthetic analysis.145 They used a sequence-to-sequence (Seq2Seq) model that used encoder-decoder-based natural NLP transformers to map a SMILES representation of the reactants to a SMILES representation of the respective products and vice versa. The NN architecture employed uses bidirectional LSTM
cells with an additive attention mechanism for token-wise alignment. This model was shown to be comparable to rule-based expert systems in solving retrosynthetic reaction prediction tasks. Other template-free approaches that have also been reported to show promising results include the use of a graph, chemical reaction networks, and similarity-based methods. Some popular retrosynthetic planning tools include AiZynthFinder, Spaya.ai (https://spaya.ai), and the Chemistry42™ platform.

The first report describing the successful execution of a multi-step synthesis route proposed by a synthesis planning software was disclosed in 2018 by Kluczniak et al. The authors used the Synthia™ software to design synthetic pathways for eight structurally diverse and synthetically challenging target molecules (Fig. 7). Synthia™, which relies on a library of ~50,000 expert encoded reaction rules compiled over a period of 15 years, was able to propose synthetic routes within 15–20 min for all targets. Interestingly, the synthetic routes proposed by Synthia™ were significantly different from their original synthetic routes disclosed in the patents and provided higher yields in fewer synthetic steps.

Additionally, synthesis planning tools have opened the possibility of enumerating and exploring the synthetically accessible chemical space. An early example is the creation of the reaction-driven Pfizer Global Virtual Library (PGVL). Enumeration of a synthetically accessible chemical library involves the use of reaction-based enumeration tools that use well-established reactions and exclusion rules, proprietary chemical ‘know-how’ protocols, and building block availability information. The Enamine REAL Space library (https://enamine.net) is one prime example of a ‘make-on-demand’ library, comprising ~15.5 billion compounds at the last count, making it the largest commercially available library of its kind. Although AI has shown great promise in streamlining synthetic organic chemistry, there are still opportunities for further improvements, such as the reliable prediction of stereochemical outcomes, reagent prediction, reaction conditions, and so forth.

**Challenges in the application of AI to preclinical drug discovery**

AI has generated a wave of excitement and investment within the biopharmaceutical industry. Although proponents of AI technology believe that it will usher in a new era of AI-driven drug discovery, skeptics argue that most of the promises are tantalizing and aspirational. However, most experts agree the reality...
clinical drug discovery is riding the crest of an era of AI-driven drug discovery. Currently, AI in preclinical drug discovery is riding the ‘peak of inflated expectations’ phase of the of Gartner’s hype cycle. Hence, it is essential to sift hyperbole from reality and set realistic expectations.

There are many challenges in implementing AI into drug discovery, one of the most demanding of which is the requirement of large amounts of high-quality training data. Building a useful and predictive model for decision-making largely depends on both the quality of the data and data set size. Unfortunately, data generation in drug discovery is both resource intensive and time consuming, often leading to a compound profiling strategy that measures few endpoints during the early stages that covers many project compounds, followed by intensive profiling of a small number of compounds during the late stages to support progression. This is evident from the millions of bioactivity data points available in commercial and public databases, such as GOSTAR, PubChem, and ChEMBL, but relatively fewer associated ADME data points, leaving us with an incomplete data matrix. Recent advances, such as sparse AI methods and deep imputation methods, could help mitigate data paucity issues.

Standardization and integration of the available drug discovery data also present a challenge during data curation. Assay readouts are often expressed in different formats (e.g., IC_{50}, EC_{50}, K_{d}, K_{i}, or % inhibition) that are not readily compatible, and the underlying data types can be discrete or continuous. Furthermore, the readouts are assay specific and comparable only under certain conditions, because they could differ in the assay format, protein construct length, substrate concentration, and so on. Hence, integrating and standardizing both public and proprietary data to expand the usable data volume is a significant challenge.

Uncertainty in drug discovery data is pervasive, and an estimate of the experimental uncertainty of K_{i} values for compounds with multiple activity values reported in ChEMBL had revealed a mean unsigned error (MUE) of 0.44 pKi units. The inherent noise associated with experimental data is referred to as aleatoric uncertainty, and the uncertainty value in the training data sets the upper limit of performance that a predictor model can achieve. In addition, drug discovery data often span a small dynamic range (2–3 log units), which often limits model predictivity.

The chemical space encompassed by an ML/DL model is termed the ‘applicability domain’ (AD) and this space is necessarily minuscule compared with the available chemical space, estimated to be ~ 10^{60} molecules. In this way, every conceivable global model is a local model, and predictions carried for compounds outside the AD of the model are generalizations based on inductive inference, which increases the uncertainty in the predictions for such compounds.

Another limitation of using ML and DL algorithms is the lack of transparency, because they operate as ‘black boxes’, implying that the features, functions, and weights encoded by these models are beyond the interpretation of a human user. The opaque decision-making process used by these algorithms cannot help discovery scientists make prospective suggestions. Thus, such opaque models are used to predict the properties of compounds already designed, limiting the utility of the models to support decision-making. Thus, for compound ideation, explainable AI (XAI), which provides transparent, informative, and interpretable findings to drive compound ideation, is needed. However, modern XAI algorithms are now emerging that assist with the interpretation of these black-box models. These algorithms use approaches such as sensitivity analysis, variable importance, and partial derivatives to extract those variables or substructural features used by the model for prediction.

Preclinical data rely on using proxy measures, such as cellular target engagement as a predictor of human in vivo target occupancy, patient-derived xenograft (PDX) mouse tumor models as a proxy for clinical efficacy in oncology programs, human HepG2 cells as a surrogate for genotoxicity, and the Caco-2 cell permeability assay as a surrogate for estimating human intestinal permeability. Although preclinical data help support clinical translation, these surrogate or proxy in vivo data points cannot be reliably used to train AI models to predict clinical outcomes (e.g., human PK, clinical efficacy, safety, and tolerability).

Apart from the scientific and technical challenges stated above, organizational culture and agility are also crucial for the adoption and implementation of AI. There is certainly some level of entrenchment and a greater degree of risk aversion in implementing new technologies in many organizations. Implementing AI technologies, to stay ahead, while realizing there is no guaranteed success, would require making some bold decisions and vision from senior leadership.

Concluding remarks
The application of AI technologies holds great promise for bringing down drug discovery timelines and cost. Although AI might not be a panacea for every problem in drug discovery, it is clearly a valuable tool if applied in the correct context and with the right data. The strength of AI technologies will certainly be used to complement human intelligence and augment our capabilities, perhaps changing the way we approach drug discovery, but not as a replacement for human ingenuity. Drawing parallels between Google’s DeepMind besting a human professional Go player with drug discovery is incongruous; drug discovery is high-dimensional science displaying nonlinearity with many known unknowns and unknown unknowns, not a Go game that can be defined based on a finite set of rules. Although one needs to be receptive in embracing new technologies such as AI for drug discovery, healthy skepticism and caution are advisable as the field matures.

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that might be perceived as affecting the perception of this review.

Acknowledgments
The authors are grateful to Drs. Sreetama Das, Jun Pei, and Vigneshwaran Namasivayam for their valuable comments and discussions. Disclaimer: Views expressed in this article are...
soley those of the authors and do not necessarily reflect the view of affiliated institutions.

References