ELSEVIER



# Machine Learning guided early drug discovery of small molecules

**Nikhil Pillai [a,*,1], Aparajita Dasgupta [b,1], Sirimas Sudsakorn [c], Jennifer Fretland [c], Panteleimon D. Mavroudis [a]**

[a] Quantitative Pharmacology, DMPK, Sanofi US, Waltham, MA, USA
[b] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[c] DMPK, Sanofi US, Waltham, MA, USA

**Machine learning (ML) approaches have been widely adopted within the early stages of the drug discovery process, particularly within the context of small-molecule drug candidates. Despite this, the use of ML is still limited in the pharmacokinetic/pharmacodynamic (PK/PD) application space. Here, we describe recent progress and the role of ML used in preclinical drug discovery. We summarize the advances and current strategies used to predict ADME (absorption, distribution, metabolism and, excretion) properties of small molecules based on their structures, and predict structures based on the desired properties for molecular screening and optimization. Finally, we discuss the use of ML to predict PK to rank the ability of drug candidates to achieve appropriate exposures and hence provide important insights into safety and efficacy.**

**Keywords: Drug discovery; Small molecules; Machine learning; Candidate selection; Molecular screening**

## Introduction

Drug discovery is a complex and challenging task and diagnostic analysis has indicated that the efficiency of R&D continues to decline.[1] The current paradigm in drug discovery is a well-defined process that moves from 'target identification' to 'lead identification' when compounds show activity against a biological target, to 'lead optimization' when ADME and potency are optimized, and finally to 'candidate selection' when a clinical candidate is chosen to advance into safety studies. This entire process can be time-consuming and resource-intensive with high reliance on translational approaches that involve assumptions that might not be possible to validate owing to lack of human data and, hence, might be incorrect for the drug studied. As a result of these assumptions, a new molecular entity (NME) could reach later stages of drug development before it is known whether it will elicit a sufficient response in humans. These late-stage failures contribute to large capital losses and higher drug development costs overall.[2]

During the lead optimization stage of discovery, molecules are evaluated using various *in vitro* assays to characterize potency, physicochemical properties and ADME properties. This is followed up with preclinical *in vivo* studies for the characterization of pharmacokinetics (PK) and pharmacodynamics (PD). Whereas PK is the study of a drug's kinetics that are largely dependent on the body's ADME processes, PD quantifies the effect of the drug in the body and it can include multiple dynamics such as biomarker response, tumor progression, cytokine release and others.[3] Several physicochemical properties of the drug affect its PK behavior including molecular weight, lipophilicity and permeability. Additionally, the physiology of the body can challenge the drug's exposure and consequently its efficacy.[4]

* *Corresponding author.Pillai, N. (nikhil.pillai@sanofi.com)*
[1] These authors have contributed equally to the manuscript.

1359-6446/
https://doi.org/10.1016/j.drudis.2022.03.017

The data generated across research are integrated into a translational approach used to predict a safe and effective clinical dose and regimen.[5–7] Although it is difficult to predict clinical efficacy based on intrinsic compound properties or *in vivo* preclinical behavior, the major factors affecting efficacy are generally attributed to the ability of a drug to reach an efficacious exposure safely.

Technical improvements in instrumentation and quantification methods have enabled large numbers of molecules to be screened for potency and ADME properties resulting in the triage of large numbers of molecules to identify a high-quality drug candidate. This has resulted in the generation of large datasets that can be used for machine learning (ML) purposes to predict various properties based on molecular structures. These large datasets can be incorporated into ML models that can decrease the risk profile of an NME in the absence of experimentation. By utilizing *in silico* ML models, one can increase the number of compounds screened as well as reduce screening times. This paradigm has enabled researchers to shift from a 'trial-and-error' approach that relies solely on expert intuition toward a more efficient and automated screening and selection strategy. Although multiple efforts have been documented at the very early stages of the drug discovery pipeline, such as in the use of target identification and hit finding, the potential for the application of these techniques toward the later stages of the process remains unclear. We believe that the application of ML can significantly reduce the experimental burden and timelines currently spent on characterization of drug response *in vitro* and *in vivo*. To this end, there is now a growing body of work that attempts to characterize and capture the implicit relationship that exists between molecular structure, properties and PK behavior.[8–10]

In this review, we attempt to analyze and reflect on this growing body of work, particularly focusing on the role ML has in the preclinical setting for decreasing uncertainty around choosing the optimal clinical candidate. We will focus on the aspects of small-molecule drug discovery within the preclinical setting where ML approaches can be utilized, especially focusing on molecular screening and optimization and candidate selection. To further aid the reader and help them contextualize the field, we provide a brief introduction to the molecular representation methods used in ML tasks in Box 1[11–22] and the

**Box 1** Molecular representation methods used in machine learning models

An important consideration when designing a machine learning (ML) algorithm for molecular property prediction or generation tasks is the molecular representation method that will be used. This can influence the architecture of the ML model used as well as the computing complexity that predetermines the computational resources needed. Here, we briefly describe the molecular representation techniques used in most ML algorithms that are designed for drug discovery and development tasks. Molecular representation approaches are classified into three categories: (i) descriptor based; (ii) natural language based; or (iii) graph-embedding based (Fig. 1). Descriptor based molecular representation is further divided into two large subsets – the first uses the quantitative properties derived from the molecular structure directly, such as those typically calculated by cheminformatics software such as RDKit.[11] These encode specific functional properties of the molecule and can be either one-dimensional (i.e., encode only a single property such as molecular weight) or multidimensional (i.e., encode multiple physicochemical aspects of the molecule as in the case of eccentricity or spherocity). The second descriptor-based approach makes use of molecular fingerprints, which are computational methods for mapping chemical space. These fingerprints encode a machine-readable representation of the structure and are generally bit vectors, and less commonly count vectors (Fig. 1a).[11,12]The second type of molecular representation is based on natural language principles which applies formal grammar rules to define the molecular structure such as SMILES (simplified molecular-input line-entry system)[13] or InChI (international chemical identifier)[14] (Fig. 1b). These representations are commonly used in deep learning algorithms for property prediction[15] as well as molecule generation.[16] An extension of these methods is SMARTS (SMILES arbitrary target specification), which specifies substructural patterns in molecules.[17] Language-based representations are also commonly used to store information in molecular databases and can later be used to convert to molecular descriptors for use in ML approaches such as random forests[18] or support vector machines.[19]Graph-based approaches are a natural adaptation of the principles of graph theory. Whereas in graph theory graphs are viewed as a collection of nodes and edges, molecules can be viewed as a collection of atoms (analogous to nodes) and bonds (analogous to edges) (Fig. 1c). Graph-based representations have gained traction in recent years. These methods are used in graph-based deep learning methods such as graph convolutional networks (GCNs)[20] or message passing neural networks (MPNNs),[21]

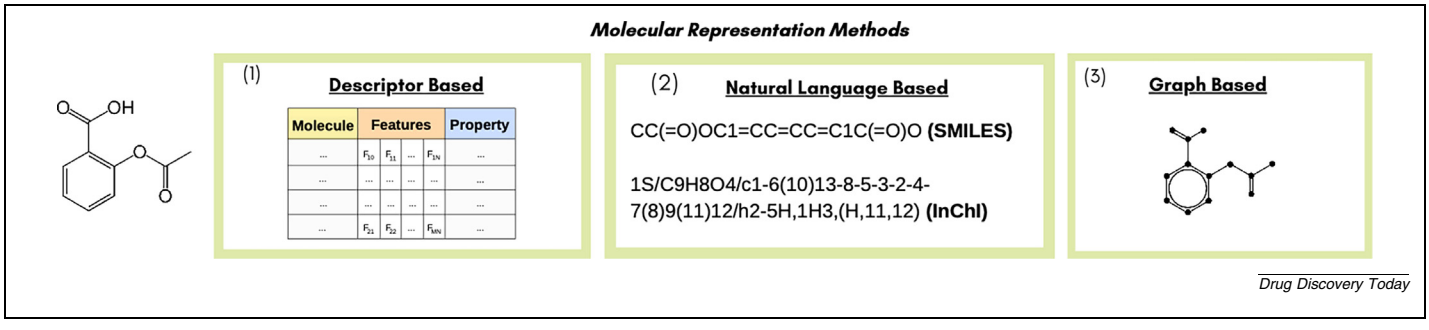


**FIG. 1**

Overview of methods used in molecular representation and machine learning models used in this domain. Here, we use the example of the commonly used nonsteroidal anti-inflammatory drug 2-acetoxybenzoic acid (generic name: aspirin) to illustrate **(a)** descriptor-based approaches, **(b)** natural-language-based approaches and **(c)** graph-based approaches.

**Box 1** (continued)
where they have demonstrated superior performance in property prediction and molecule generation tasks. However, these representations still face issues associated with representing stereoisomerism,[22] where these methods are unable to distinguish between different stereoisomers owing to their isomorphic structure.

**Box 2** An introduction to machine learning algorithms used in drug discovery and development
Here, two main groups of machine learning (ML) methods are discussed. These are grouped into traditional ML methods (such as tree-based methods, latent variable methods) and deep learning methods (Fig. 2a,b). We provide readers with a brief description of the methods mentioned in the text.The most commonly used ML methods comprise decision trees or tree-based methods.[23] These methods are non-parametric methods that are used for regression and classification. An example of tree-based algorithms is random forests where multiple random subsets of the model input variables are used to generate multiple decision trees[18] and the output is either the mean (regression) or mode (classification) of the trees generated (Fig. 2a). Another algorithm is the support vector machines (SVM), which operates on the concepts of hyperplanes that either best divides the classes investigated (for classification) or fits the maximum number of points.[19] Gaussian processes are a non-parametric Bayesian method[24] that are also very commonly used. Finally, latent variable methods such as partial least squares where the predictor and target variable methods are linearly projected while linking the relation between the two[25] are utilized in QSAR models as well.Deep learning methods are the second class of models that are utilized within the domain (Fig. 2b). Some of the deep learning methods that are used in this class of models include autoencoders – a type of neural network that learns a lower dimensional representation of the input space – and recurrent neural nets, which allow for modeling temporal relations such as in language models[26] (or molecular string representations). Other deep learning methods are reinforcement learning methods, which are a different class of models from supervised and unsupervised learning. These models aim to maximize a reward function using a trial-and-error approach within a given environment[27] such as generating molecules with optimum properties within a given design space.[28]The use of these ML models is decided on a variety of factors, including the inherent characteristics of the task, data availability and the underlying assumptions regarding model functions. In early drug discovery, where large amounts of data exist in the form of molecular candidate search space,[29,30] deep learning methods were found to be particularly favorable, owing to the data-hungry nature of these algorithms. In smaller data domain tasks such as in vivo modeling, simple machine learning tasks can be more useful because deep learning methods might not accurately capture underlying trends within these limited data regimes. The underlying assumptions regarding model parameters also play an important part in algorithm selection. For example, if the underlying response surface is assumed to be linear, partial least squares can be used. However, in the case of a nonlinear response surface, other ML models such as support vector regression, gaussian processes or tree-based methods are more appropriate.

ML algorithms mentioned in the remaining text are overviewed in Box 2.[18,19,23–30]

## Machine learning applied to molecular screening and optimization

Once a suitable target is identified and the intended biomarkers are decided upon, the next step in the drug discovery process is to find molecules that exhibit therapeutic activity against the specific target and optimize these potential molecules. This is termed as the 'hit-to-lead' generation phase.[31] At the hit-finding stage, HTS is used to identify compounds from a library of potential candidates that exhibit activity against the target. However, despite being the current state-of-the-art method for screening, HTS does not necessarily yield a viable molecule. This holds particularly true when one is reminded of the extensively large search space that entails such endeavors. For example, the ZINC database[29] contains 750 million purchasable compounds for screening.

When dealing with such large and intractable spaces, the use of ML techniques can greatly increase the number of viable hits. As an example, Mehta and colleagues[32] demonstrated the use of Bayesian learning for an efficient search of the molecular space using the docking score as the proxy, and were able to discover 90% of the top hits while performing docking for only 6% of the space. They tested three molecular representation techniques: extended-connectivity fingerprints (ECFP); Mol2Vec (an unsupervised pre-training method to generate molecular vectors); and continuous and data-driven descriptors (CDDD), for their Bayesian Optimization Framework. The surrogate ML models tested for the Bayesian Optimization Framework were the Gaussian Process and the Deep Gaussian Process. The authors tested their models on the ZINC-250 k dataset containing 250 000 molecules, the Enamine dataset containing 2.1 million molecules and the Ultra Large Docking Library, containing 96 million molecules. The virtual screening effort was tested against two protein receptors: Tau-tubulin kinase 1 (TTBK1), often targeted for neurodegenerative diseases; and the main protease of SARS-CoV-2. The authors found that the optimal method of molecular representation largely depended on the size of the dataset being interrogated. The choice of surrogate model was also dependent on the search space but to a lesser extent. Their work brings to light an important issue in current ML literature for molecular sciences, which is that molecular representation can largely influence the performance of an algorithm and hence must be designed such that the representation matches the target to be interrogated.

Another common approach that is utilized during the hit-to-lead optimization phase is the use of QSAR for property prediction. QSARs are any computational modeling method used for revealing relationships between structural properties of chemical compounds and biological activities. The use of accurate ML algorithms for this purpose can reduce the experimental burden significantly by informing medicinal chemists of the best candidate for a given target without the need for *in vitro* and *in vivo* experimentation, thus conserving time and resources. An example of such an approach is the work by Wang *et al.*[33] which used various ML algorithms to predict the human intestinal perme-

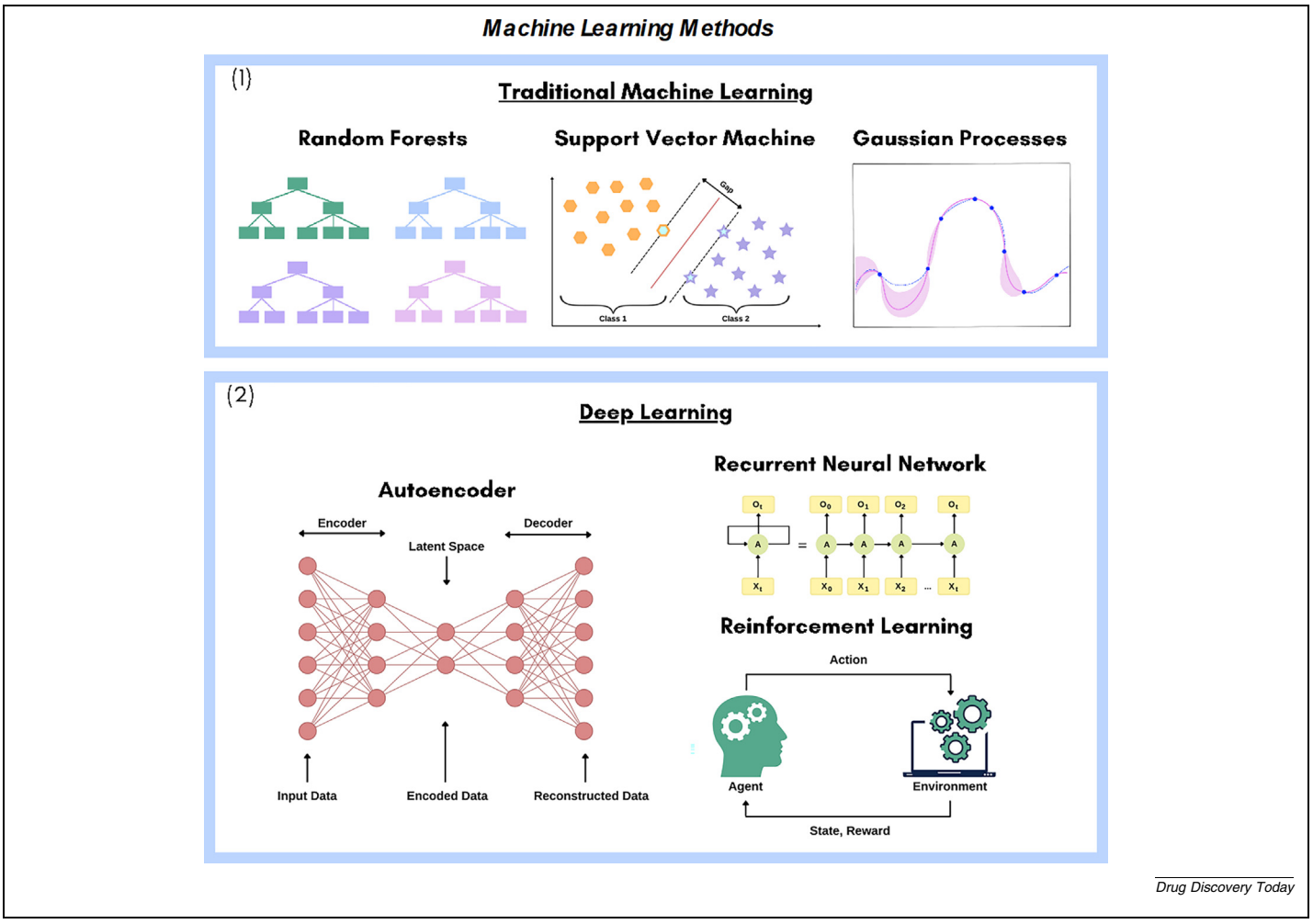


**FIG. 2**

Machine learning methods are chiefly classified into: **(a)** traditional machine learning methods, which include low-data methods such as random forests, support vector machines and gaussian processes; and **(b)** deep learning approaches which utilize neural networks.

ability using the permeability coefficient of the human adenocarcinoma cell line (Caco-2) as a cell culture model. Their work used multiple linear regression, partial least squares regression, support vector machine regression and boosting and found that boosting algorithms were most suitable for their application. The inputs to all ML models were 193 2D and 3D molecular descriptors. In addition to creating a predictive model for permeability, they were also able to ascertain which descriptors were most important using a descriptor ablation approach and were able to identify and analyze the underlying biological mechanisms that their models were able to capture. The advantages of these models can further be understood, especially when one considers the traditionally high cost and long duration of the cell culture periods (21–24 days) required for performing experiments to estimate the *in vivo* drug permeability.

QSAR models have also expanded significantly so that graphical user interface (GUI)-based web platforms can be used to interrogate potential hits for evaluation of multiple properties. ADMETlab is a platform that performs drug likeness analysis, ADME predictions, systemic evaluations and similarity searching against a large database of >280 000 entries. The user can upload their own input structures in the form of SMILES strings or SDF files or draw the structure using the online editor. Possible analyses include drug-likeness predictions using Lipinski, Ghose, Oprea, Veber and Varma rules, as well as a classification model developed using input structures from the DrugBank and CHEMBL[30] databases. The drug-likeness classification model was built using MACCS fingerprints as input and a random forests classifier as the underlying ML model. Additionally, the module has nine regression models and 22 classification models for prediction of ADMET endpoints. All models use molecular descriptors and fingerprints as input and use low data ML methods such as random forests, support vector machines and partial least squares.[34]

SwissADME is another example of web-based tools that provide models to predict drug likeness, physicochemical properties and PK behavior. Similar to ADMETlab, the web tool contains models using physicochemical descriptors and open-source fingerprints as input to evaluate bioavailability using lipophilicity, size, polarity, solubility, flexibility and saturation as representative properties. Multiple PK models including QSARs developed to predict skin permeability using the multiple linear regression suggested by Potts and Guy, the blood–brain barrier permeation and the passive human gastrointestinal absorption using the BOILED-Egg model by Daina and Zoete are also PK outputs within the SwissADME modules. Furthermore, one can assess

the synthetic accessibility, the promiscuousness (using PAINS filters) and the 'lead-likeness' using Brenk filters of the molecules tested to find optimizable candidates.[35] These models and platforms provide accessible user interfaces in which medicinal chemists can evaluate large numbers of molecules and facilitate rank ordering of compounds with reasonable confidence. There has been a significant increase in the development of QSAR models that are specific to certain targets as well as generalized evaluations of small-molecule therapeutic activity. Among notable works, Neves *et al.*[36] have succinctly summarized the use of QSAR models in virtual screening.

Complementary to the task of predicting biologically relevant properties for a given molecule is the inverse QSAR or molecule generation problem that pertains to the generation of novel compounds that possess specific properties of interest. To better appreciate the challenges of the inverse QSAR and the strategies employed thereby, an understanding of the journey to elucidate a potential candidate from the molecular search space is warranted. A systematic investigation of all possible molecules to find an optimum therapeutic candidate is an especially daunting task, primarily owing to the massive search space within the small-molecule domain. Estimates of all possible candidates within this space range between $10^{20}$ and $10^{60}$ possible molecules, which is dependent on the search criteria.[37] Recent progress in the compilation of small-molecule databases such as ZINC[29] and ChEMBL,[30] as well as advances in molecular representation, computational methods and processing capabilities, has led to significant progress and several studies that are focused on generation of random drug-like molecules as well as target-specific molecules using generative neural-network-based models.[16]

One example of such work is the MolGPT model developed by Bagal *et al.*[38] In this work, the authors take inspiration from current state-of-the-art natural-language-processing models, which are generative pre-training (GPT) models to train a transformer-based architecture using masked self-attention to predict a sequence of SMILES strings for molecular generation. They showed that their models were able to represent the chemical space with high accuracy, in terms of molecular diversity, validity and specific properties such as the topological polar surface area (TPSA) and partition coefficients such as log*P* among many others. Another approach that is used for this class of problems is reinforcement learning. Popova *et al.* utilized generative and predictive neural networks to generate chemically feasible SMILES strings with desired physicochemical and physiological properties.[39] In this study, the authors used SMILES strings as input molecular representations where a generative model and a predictive model are trained separately using a supervised learning approach, and then trained jointly within a reinforcement learning setting to generate novel molecules that are geared toward a specific physicochemical property such as melting point or hydrophobicity, as well as similar compounds with inhibitory activity against targets such as Janus protein kinase 2.

In addition to the generation of novel therapeutic molecules, one must also consider feasibility constraints such as the 'synthesizability' of suggested compounds to ensure a commercially viable product. Retrosynthesis analysis, which is the process of finding suitable starting materials to produce a given molecule, is an important consideration in molecule generation tasks. A growing body of work has now contributed to a state where there are now massive databases and search algorithms that are not only able to identify synthesizable molecules with properties of interest but can also suggest new, more efficient routes of synthesis.[40] These approaches can be grouped into two main categories. Template-based approaches use neural networks to rank possible chemical reaction pathways that are previously manually encoded or derived from databases.[41] Template-free approaches, by contrast, do not make any assumptions on possible reaction pathways and use graph-based[42] or natural-language-based[43] approaches to predict reaction pathways from commercially available starting materials to the molecules of interest. These studies are an important step toward the identification and design of drug candidates that not only possess the intended medicinal properties but are practically synthesizable as well.

## Hybrid approaches for candidate selection

In addition to molecular property prediction and inverse molecular generation, the prediction of human PK and PD response at the point of design are also of immense importance in identifying the most likely candidate that can provide sufficient drug exposure to elicit the desired pharmacological effect in the clinic.[44] An important first step toward this goal is to predict PK parameters such as area under the drug's concentration vs time curve (AUC), clearance (CL), volume of distribution ($V_d$), half-life ($t_{1/2}$), maximum concentration ($C_{max}$), time to reach $C_{max}$ ($t_{max}$) and bioavailability (F). These are quantities emanating from the drug's concentration vs time curve and give an overall indication of the behavior of the drug in the body. To date, there are several studies of *in silico* modeling that can predict human and animal PK parameters from chemical structures. Of note, Kosugi and Hosea showed that ML models, particularly the random forests model and radial basis function model (among eight ML models tested) for CL prediction in rats provided a good alternative to traditional approaches (e.g., *in-vitro – in-vivo* extrapolation), with a potential to be used earlier in the drug discovery pipeline.[45] The input to these models consisted of 330 2D-SMARTS-based descriptors. Although these models obtained comparable performance, in most cases they were limited in predicting only one or two PK parameters and more importantly owing to their black-box nature did not incorporate a mechanistic understanding of the underlying ADME processes which are determinants of PK parameters.

To address the limitations embedded in the black-box architecture of ML models, hybrid approaches that use a combination of ML and mechanistic modeling are being investigated.[6,7,46] In particular, these approaches incorporate the ML-driven compound-related information such as ADME properties (permeability, pKA, lipophilicity, intrinsic metabolism, etc.) with the physiologically based pharmacokinetic (PBPK) model framework that uses a set of ordinary differential equations (ODE) and physiological parameters such as blood flow, tissue volumes, hematocrit and metabolizing enzyme expression to describe the transport of drug via blood in the different tissue compartment within the body to predict the PK profile. Hosea and Jones demonstrated that *in silico* and *in vitro* information can be utilized to predict PK profiles by employing commercially available

software packages such as ADMET predictor and Gastroplus.[46] Antontsev *et al.* used a hybrid approach that combines ML optimization with mechanistic modeling to simulate compound plasma concentration profiles[47] and were able to demonstrate with high accuracy the drug concentration–time profile and tissue partition coefficient while varying log*P* (lipophilicity descriptor). Finally, Chen *et al.* have used a hybrid approach to develop decision-tree-based methods, which incorporates mechanistic PK/PD approaches within the training dataset.[48] This approach trained the model on a smaller dataset of known compounds to uncover the relationship that exists between the ADME parameters and the output parameters of a PKPD or PBPK model for a larger sample of unknown compounds. These models not only have specific utility within the early drug discovery pipeline, primarily owing to the small number of data points that exist in this space, but can also help identify relationships between the concentration–time profiles and predicted ADME properties despite the data being limited.[48]

In addition to predicting the PK profile from ADME properties, the translation of PK and PD response profiles from the preclinical to clinical phase plays an important part in selecting drug candidates. Although the traditional PK/PD approaches have proven to be robust for dose predictions, their implementation remains a challenge owing to the lack of data and validated translational approaches from preclinical models to humans. This results in models developed in early drug discovery having to incorporate many assumptions where validation is challenging.[49] As a result, there has been a growing interest in the utilization of ML techniques to provide more-robust predictions of parameters used in the translational modeling efforts. Lu *et al.* developed a neural PK/PD model that predicts dose–response curves that appear to be generalizable and applicable to untested dosing regimens.[50] Similarly, Kosugi *et al.* compared a mechanistic neuropharmacokinetic (neuroPK) model to two ML approaches (random forests regression and gaussian processes) to predict unbound brain-to-plasma partitioning. Their analysis concluded that ML models performed better when compared with the neuroPK model within the chemical applicability domain but performed worse when an external testset was used.[51] This highlighted a chief drawback in ML models which is the loss in predictive capability when extrapolating beyond the training dataset used, specifically in the case of tree-based methods such as random forests. Nonetheless, the current ML approaches are very useful in cases where significant knowledge gaps exist such as the prediction of bioavailability where values are unpredictable and highly variable[52] or in the area of personalized medicine where it might be challenging to use traditional population PK methods to distinguish between individual patients and decide on individual dosing strategies.[53,54]

## Concluding remarks

The use of ML approaches in early drug discovery has been gaining traction in recent years. Improvements in molecular representations, computational approaches and computing capabilities have contributed to significant improvement of these methods and more-widespread adoption, particularly within the context of early drug discovery. In this paper, we highlight the recent advances within these fields with a focus on molecular screening and optimization, and candidate selection. There is growing evidence that utilization of ADMET prediction using QSAR models, inverse molecular generation and PK prediction can be of the utmost importance for molecular screening and optimization. Furthermore, application of ML models along with mechanistic models can be used for molecular screening and optimization which can ultimately help in rank ordering and elucidating top candidates. The chief advantages of applying ML to these domains include increased sampling efficiency, reduced experimental burdens and timelines and improved identification of safe and efficacious molecules.

Despite this, challenges remain within the field including those relating to preclinical and clinical data availability, representation and uniformity. Data standardization represents a large opportunity within the field to encourage cross collaboration between different stages of the drug discovery and development pipeline. Additionally, we need to ensure improved communication across the stages of drug development and a recursive approach where corrective actions can be undertaken in real time to prevent nonefficacious molecules from moving into later stages of clinical development. These challenges need to be addressed to promote the adoption of these modeling frameworks across the pharmaceutical industry.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


1 J.W. Scannell, A. Blanckley, H. Boldon, B. Warrington, Diagnosing the decline in pharmaceutical R&D efficiency, Nat Rev Drug Discov 11 (2012) 191–200, https://doi.org/10.1038/nrd3681.

2 A. Schuhmacher, O. Gassmann, M. Hinder, Changing R&D models in research-based pharmaceutical companies, J Transl Med 14 (2016) 105, https://doi.org/10.1186/s12967-016-0838-4.

3 D.W. Nebert, G. Zhang, Pharmacogenomics, in: R.E. Pyeritz, B.R. Korf, W.W. Grody (Eds.), *Emery and Rimoin's principles and practice of medical genetics and genomics*, 7th ed., Academic Press, 2019, pp. 445–486, https://doi.org/10.1016/B978-0-12-812537-3.00016-0.

4 T. Velkov, P.J. Bergen, J. Lora-Tamayo, C.B. Landersdorfer, J. Li, PK/PD models in antibacterial development, Curr Opin Microbiol 16 (2013) 573–579, https://doi.org/10.1016/j.mib.2013.06.010.

5 D.B. Kassel, Applications of high-throughput ADME in drug discovery, Curr Opin Chem Biol 8 (2004) 339–345, https://doi.org/10.1016/j.cbpa.2004.04.015.

6 D. Cook, D. Brown, R. Alexander, et al., Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework, Nat Rev Drug Discov 13 (2014) 419–431, https://doi.org/10.1038/nrd4309.

7 P. Morgan, D.G. Brown, S. Lennard, et al., Impact of a five-dimensional framework on R&D productivity at AstraZeneca, Nat Rev Drug Discov 17 (2018) 167–181, https://doi.org/10.1038/nrd.2017.244.

8 C. Réda, E. Kaufmann, A. Delahaye-Duriez, Machine learning applications in drug development, Comput Struct Biotechnol J 18 (2020) 241–252, https://doi.org/10.1016/j.csbj.2019.12.006.

9 J. Vamathevan, D. Clark, P. Czodrowski, et al., Applications of machine learning in drug discovery and development, Nat Rev Drug Discov 18 (2019) 463–477, https://doi.org/10.1038/s41573-019-0024-5.

10 Y. Sakiyama, The use of machine learning and nonlinear statistical tools for ADME prediction, Expert Opin Drug Metab Toxicol 5 (2009) 149–169, https://doi.org/10.1517/17425250902753261.

11 Landrum G. RDKit. Published 2010. http://www.rdkit.org/ [accessed June 29, 2021].

12 Landrum G. Fingerprints in the RDKit. Presented at: RDKit UGM 2012. London; 2012.

13 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J Chem Inf Comput Sci 28 (1988) 31–36, https://doi.org/10.1021/ci00057a005.

14 S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier, J Cheminform 7 (2015) 23, https://doi.org/10.1186/s13321-015-0068-4.

15 Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model*. Published online 2019:19.

16 O. Prykhodko, S.V. Johansson, P.C. Kotsias, et al., A *de novo* molecular generation method using latent vector based generative adversarial network, J Cheminform 11 (2019) 74, https://doi.org/10.1186/s13321-019-0397-9.

17 Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html [accessed December 19, 2021].

18 A. Liaw, M. Wiener, Classification and Regression by RandomForest, Forest 23 (2001).

19 W.S. Noble, What is a support vector machine?, Nat Biotechnol 24 (2006) 1565–1567, https://doiorg/10.1038/nbt1206-1565.

20 V. Korolev, A. Mitrofanov, A. Korotcov, V. Tkachenko, Graph convolutional neural networks as "general-purpose" property predictors: the universality and limits of applicability, J Chem Inf Model 60 (2020) 22–28, https://doi.org/10.1021/acs.jcim.9b00587.

21 Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for Quantum chemistry. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. JMLR.org; 2017:1263–72.

22 Pattanaik L, Ganea OE, Coley I, Jensen KF, Green WH, Coley CW. Message Passing Networks for Molecules with Tetrahedral Chirality. *ArXiv201200094 Cs Q-Bio*. Published online December 4, 2020. http://arxiv.org/abs/2012.00094 [accessed June 29, 2021].

23 P. Bühlmann, Bagging, Boosting and ensemble methods, in: J.E. Gentle, W.K. Härdle, Y. Mori (Eds.), Handbook of Computational Statistics: Concepts and Methods. Springer Handbooks of Computational Statistics, Springer, 2012, pp. 985–1022, https://doi.org/10.1007/978-3-642-21551-3_33.

24 C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2005.

25 P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal Chim Acta 185 (1986) 1–17, https://doi.org/10.1016/0003-2670(86)80028-9.

26 D.E. Rumelhart, G.E. Hinton, R.J. Williams, *Learning Internal Representations by Error Propagation*, California Univ San Diego La Jolla Inst Cognitive Science (1985).

27 Lillicrap TP, Hunt JJ, Pritzel A, et al. Continuous control with deep reinforcement learning. *ArXiv Prepr ArXiv150902971*. Published online 2015.

28 Simm GNC, Pinsler R, Hernández-Lobato JM. Reinforcement Learning for Molecular Design Guided by Quantum Mechanics. *ArXiv200207717 Cs Stat*. Published online June 29, 2020. http://arxiv.org/abs/2002.07717 [accessed June 30, 2021].

29 J.J. Irwin, B.K. Shoichet, ZINC – A Free Database of Commercially Available Compounds for Virtual Screening, J Chem Inf Model 45 (2005) 177–182, https://doi.org/10.1021/ci049714.

30 D. Mendez, A. Gaulton, A.P. Bento, et al., ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Res 47 (2019) D930–D940, https://doi.org/10.1093/nar/gky1075.

31 J. Hughes, S. Rees, S. Kalindjian, K. Philpott, Principles of early drug discovery, Br J Pharmacol 162 (2011) 1239–1249, https://doi.org/10.1111/j.1476-5381.2010.01127.x.

32 S. Mehta, S. Laghuvarapu, Y. Pathak, A. Sethi, M. Alvala, P.U. Deva, MEMES: Machine learning framework for Enhanced MolEcular Screening, Chem Sci 12 (2021) 11710–11721, https://doi.org/10.1039/D1SC02783B.

33 N.N. Wang, J. Dong, Y.H. Deng, et al., ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting, J Chem Inf Model 56 (2016) 763–773, https://doi.org/10.1021/acs.jcim.5b00642.

34 J. Dong, N.N. Wang, Z.J. Yao, et al., ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database, J Cheminform 10 (2018) 29, https://doi.org/10.1186/s13321-018-0283-x.

35 A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci Rep 7 (2017) 42717, https://doi.org/10.1038/srep42717.

36 B.J. Neves, R.C. Braga, C.C. Melo-Filho, J.T. Moreira-Filho, E.N. Muratov, C.H. Andrade, QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery, Front Pharmacol 9 (2018), https://doi.org/10.3389/fphar.2018.01275.

37 J.L. Reymond, M. Awale, Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database, ACS Chem Neurosci 3 (2012) 649–657, https://doi.org/10.1021/cn3000422.

38 V. Bagal, R. Aggarwal, P.K. Vinod, U.D. Priyakumar, MolGPT: Molecular Generation Using a Transformer-Decoder Model, J Chem Inf Model (2021).

39 M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design, Sci Adv 4 (2018) eaap7885, https://doi.org/10.1126/sciadv.aap7885.

40 S. Genheden, A. Thakkar, V. Chadimová, J.L. Reymond, O. Engkvist, E. Bjerrum, AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning, J Cheminform 12 (2020) 70, https://doi.org/10.1186/s13321-020-00472-1.

41 M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (2018) 604–610, https://doi.org/10.1038/nature25978.

42 C. Shi, M. Xu, H. Guo, M. Zhang, J. Tang, A Graph to Graphs Framework for Retrosynthesis Prediction abs/2003.12725, CoRR (2020). https://arxiv.org/abs/2003.12725.

43 S. Zheng, J. Rao, Z. Zhang, J. Xu, Y. Yang, Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks, J Chem Inf Model 60 (2020) 47–55, https://doi.org/10.1021/acs.jcim.9b00949.

44 N.H.G. Holford, H.C. Kimko, J.P.R. Monteleone, C.C. Peck, Simulation of Clinical Trials, Annu Rev Pharmacol Toxicol 40 (2000) 209–234, https://doi.org/10.1146/annurev.pharmtox.40.1.209.

45 Y. Kosugi, N. Hosea, Direct Comparison of Total Clearance Prediction: Computational Machine Learning Model versus Bottom-Up Approach Using *In Vitro* Assay, Mol Pharm 17 (2020) 2299–2309, https://doi.org/10.1021/acs.molpharmaceut.9b01294.

46 N.A. Hosea, H.M. Jones, Predicting Pharmacokinetic Profiles Using *in Silico* Derived Parameters, Mol Pharm 10 (2013) 1207–1215, https://doi.org/10.1021/mp300482w.

47 V. Antontsev, A. Jagarapu, Y. Bundey, et al., A hybrid modeling approach for assessing mechanistic models of small molecule partitioning *in vivo* using a machine learning-integrated modeling platform, Sci Rep 11 (2021) 11143, https://doi.org/10.1038/s41598-021-90637-1.

48 E.P. Chen, R.W. Bondi, P.J. Michalski, Model-based Target Pharmacology Assessment (mTPA): An Approach Using PBPK/PD Modeling and Machine Learning to Design Medicinal Chemistry and DMPK Strategies in Early Drug Discovery, J Med Chem 64 (2021) 3185–3196, https://doi.org/10.1021/acs.jmedchem.0c02033.

49 H. Zou, P. Banerjee, S.S.Y. Leung, X. Yan, Application of Pharmacokinetic-Pharmacodynamic Modeling in Drug Delivery: Development and Challenges, *Front Pharmacol* 11 (2020), https://doi.org/10.3389/fphar.2020.00997.

50 J. Lu, B. Bender, J.Y. Jin, Y. Guan, Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling, Nat Mach Intell 1–9 (2021), https://doi.org/10.1038/s42256-021-00357-4.

51 Y. Kosugi, K. Mizuno, C. Santos, S. Sato, N. Hosea, M. Zientek, Direct Comparison of the Prediction of the Unbound Brain-to-Plasma Partitioning Utilizing Machine Learning Approach and Mechanistic Neuropharmacokinetic Model, AAPS J 23 (2021) 72, https://doi.org/10.1208/s12248-021-00604-x.

52 H. Lou, M.J. Hageman, Machine Learning Attempts for Predicting Human Subcutaneous Bioavailability of Monoclonal Antibodies, Pharm Res 38 (2021) 451–460, https://doi.org/10.1007/s11095-021-03022-y.

53 You W, Widmer N, De Micheli G. Personalized modeling for drug concentration prediction using Support Vector Machine. In: *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*. Vol. 3; 2011. p. 1505–9. doi: 10.1109/BMEI.2011.6098593.

54 T. Koizumi, T. Suzuki, N.S. Pillai, et al., Circadian patterns of hallucinatory experiences in patients with schizophrenia: Potentials for chrono-pharmacology, J Psychiatr Res 117 (2019) 1–6, https://doi.org/10.1016/j.jpsychires.2019.06.019.