molecular pharmaceutics

pubs.acs.org/molecularpharmaceutics

Article

Prediction of In Vivo Pharmacokinetic Parameters and Time–Exposure Curves in Rats Using Machine Learning from the Chemical Structure

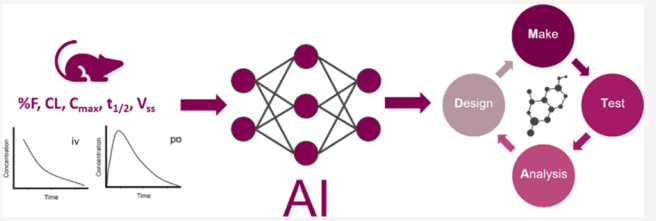
Olga Obrezanova,* Anton Martinsson, Tom Whitehead, Samar Mahmoud, Andreas Bender, Filip Miljković, Piotr Grabowski, Ben Irwin, Ioana Oprisiu, Gareth Conduit, Matthew Segall, Graham F. Smith, Beth Williamson, Susanne Winiwarter, and Nigel Greene

Cite This: https://doi.org/10.1021/acs.molpharmaceut.2c00027

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Animal pharmacokinetic (PK) data as well as human and animal in vitro systems are utilized in drug discovery to define the rate and route of drug elimination. Accurate prediction and mechanistic understanding of drug clearance and disposition in animals provide a degree of confidence for extrapolation to humans. In addition, prediction of in vivo properties can be used to improve design during drug discovery, help select compounds with better properties, and reduce the number of in vivo experiments. In this study, we generated machine learning models able to predict rat in vivo PK parameters and concentration–time PK profiles based on the molecular chemical structure and either measured or predicted in vitro parameters. The models were trained on internal in vivo rat PK data for over 3000 diverse compounds from multiple projects and therapeutic areas, and the predicted endpoints include clearance and oral bioavailability. We compared the performance of various traditional machine learning algorithms and deep learning approaches, including graph convolutional neural networks. The best models for PK parameters achieved R^2 = 0.63 [root mean squared error (RMSE) = 0.26] for clearance and R^2 = 0.55 (RMSE = 0.46) for bioavailability. The models provide a fast and cost-efficient way to guide the design of molecules with optimal PK profiles, to enable the prediction of virtual compounds at the point of design, and to drive prioritization of compounds for in vivo assays.



KEYWORDS: *rat pharmacokinetics, clearance, bioavailability, concentration–time pharmacokinetic profiles, machine learning, neural networks, data imputation, graph convolutions, QSPR, compound design*

1. INTRODUCTION

The efficacy and safety of a drug is a function of both its intrinsic molecular properties (such as bioactivity against molecular targets, chemical reactivity, etc.) and its concentration at a particular site of action as a function of time, that is, its pharmacokinetic (PK) profile.[1] While the former has received significant attention recently in the context of "artificial intelligence" (AI) in drug discovery in areas such as bioactivity prediction[2,3] and the de novo design of ligands for particular proteins,[4] the impact of AI in the area of modeling in vivo properties, such as PK, is much less pronounced at this stage. One reason is that domains differ significantly with respect to the quantity of high-quality data available.[5,6] In some areas, in vitro assays can be run to characterize compounds,[7] such as biochemical assays or assays for PK-related properties such as LogD (base-10 logarithm of distribution coefficient) or solubility, which give rise to large numbers of available data points. This makes these properties relatively amenable to current developments in the machine learning domain, such as deep learning.[8] However, in vivo PK data (as well as in vivo data more generally) are more costly and complex, resulting in a lack of data in this domain, which hinders the application of some algorithms.[6] On the other hand, due to the direct therapeutic relevance of in vivo assays, as well as their high cost, modeling these types of endpoints provides a stronger incentive to generate in silico models. Furthermore, it has been shown that failure rates in the clinical phases due to sub-optimal PK are what makes drug discovery so costly.[5,9]

In vivo rat PK studies commonly assess intravenous (iv) and oral (po) administration of a test article and measure the plasma concentration of the compound over time, typically over a 24 h period. From this concentration versus time curve, a number of PK parameters can be calculated including the area under the concentration–time curve (AUC), the

Received: January 12, 2022
Revised: March 28, 2022
Accepted: March 29, 2022

maximum plasma concentration (C_{max}), half-life ($t_{1/2}$), clearance (CL), volume of distribution (V_{ss}), and oral bioavailability (F). Drug discovery projects strive to achieve optimal PK parameters to reach efficacy in vivo in combination with a suitable safety profile at a given dosing regimen.[1] For example, for an oral drug, bioavailability needs to be high enough (e.g., $F > 30\%$) to reduce interindividual variability, while clearance needs to be sufficiently low to achieve a long enough exposure at the site of action for therapeutic efficacy.

Current approaches applied to predict in vivo PK include (among others) in vitro to in vivo extrapolation (IVIVE)[10,11] and physiologically based pharmacokinetics (PBPK) modeling.[12] The well-stirred model (WSM),[13−15] an IVIVE approach for clearance, assumes that the drug concentration in the liver is uniform ("well stirred") and estimates only hepatic metabolic clearance. The model incorporates measured in vitro data from liver microsomes or hepatocytes and protein binding and a subsequent extrapolation step to in vivo including hepatic blood flow.[14]

PBPK models are usually applied later in drug discovery projects and require a comprehensive suite of in vitro and usually also animal in vivo characterization of the compound in question. The method can be used to understand the in vivo behavior and to extrapolate it to humans. PBPK defines a compound's overall PK by describing its distribution in relevant organs (compartments), based on pharmacological parameters such as organ size, blood flow, and tissue composition in relation to the compound properties. The approach can be used to model the compound concentration in different organs as well as in different disease settings or populations.[12] However, it is not possible to readily apply this methodology in a high-throughput manner.

Significant advances have been made in simulating human in vivo PK from in vitro and preclinical in vivo data, and recent work from AstraZeneca[16] describes that "83% of AstraZeneca drug development projects progress in the clinic with no PK issues; and 71% of key PK parameter predictions [64% of area under the curve (AUC) predictions; 78% of maximum concentration (C_{max}) predictions; and 70% of half-life predictions] are accurate to within 2-fold".

For practical purposes, the direct prediction of PK parameters based on the chemical structure is desirable since it may enable drug discovery scientists to move compound prioritization from proxy properties, such as a series of in vitro properties, to the more relevant in vivo space.[5,6]

Several studies used machine learning approaches to predict human and in vivo animal PK parameters.[17−24] Wang et al.[17] established quantitative structure−property relationship (QSPR) models for four human PK parameters, including volume of distribution at the steady state, clearance, half-life, and fraction unbound in plasma, using a data set consisting of 1352 drugs (which is currently also the largest publicly available data set of its type[25]). For clearance, the model accuracy is better than for in vivo clearance models by other groups, and this might be due to the fact that iv data were modeled in this work, due to a bias toward compounds with low clearance, and due to the way the data set was derived. Also, more specific models for volume of distribution have been described recently,[18] based on random forest methods, and evaluated using an independent test set of 213 compounds, which was found to compare favorably to methods based on in vitro properties. Recently, machine learning models, predicting from the chemical structure and dose, for human PK parameters, C_{max}, AUC and volume of distribution (V_d), were built on a set of 1000 clinical compounds and further validated on AstraZeneca clinical data.[19]

Recently, deep learning and graph convolutional algorithms have been applied to in vivo PK modeling. In a study on a large data set of ∼1900 in vivo data points[20] researchers at Bayer modeled iv and oral drug exposure and oral bioavailability in rats using a variety of hybrid modeling approaches, including deep neural networks, linear mapping, and PBPK models. Compounds were described as either (a) six experimentally determined in vitro and physicochemical properties, namely, membrane permeation, free fraction, metabolic stability, solubility, pK_a value, and lipophilicity; (b) the outputs of six in silico absorption, distribution, metabolism, and excretion (ADME) models trained on the same properties; or (c) the chemical structure encoded as fingerprints or simplified molecular input line entry system (SMILES) strings. The authors found that exposure after iv administration can be predicted similarly well using experimental and predicted properties as the input. The model errors for exposure after po administration were generally higher, and the prediction from in vitro inputs performs significantly better in comparison to their in silico counterparts, which the authors attributed to the higher complexity of oral bioavailability. Using graph convolutional networks on data sets from Merck, the authors of another study[21] were able to show that their method, PotentialNet, achieves a 64% average improvement and a 52% median improvement in R^2 over random forests across all 31 data sets used in the study (which comprise a wide range of mostly ADME-related endpoints plus in vivo dog and rat PK endpoints). For in vivo endpoints, such as rat and dog clearance data, only marginal improvements in performance were seen. Using transfer learning and multitask learning,[22] one recent model was pretrained on over 30 million bioactivity data points, and then, four human PK parameters for 1104 FDA-approved small-molecule drugs were modeled, namely, oral bioavailability, plasma protein binding, apparent volume of distribution at the steady state, and elimination half-life. The multitask learning model generally has shown the best performance for the endpoints modeled, although not with a very large margin in some cases.

One key question is whether machine learning models for in vivo PK properties perform better than extrapolating from in vitro data using mechanistic IVIVE approaches. Kosugi and Hosea compared IVIVE and machine learning approaches for in vivo clearance prediction in rat[23] on a structurally diverse set of 1114 compounds with known in vitro intrinsic clearance and plasma protein binding. The predictivity of machine learning models was generally improved by incorporating in vitro parameters as input features. On the other hand, clearance prediction utilizing in vitro intrinsic clearance data in combination with the WSM was found to perform substantially worse compared to machine learning approaches. Similar conclusions were obtained in a study by the same authors, which compared machine learning models for the in vivo AUC after po administration to the IVIVE approach using a data set of 595 compounds.[24] Both of these studies, in agreement with our findings in the current work, suggest that

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

in silico machine learning models for compound in vivo PK properties are of practical value in drug design.

As exemplified above, there exists a prior art in the area of modeling in vivo PK parameters based on the chemical structure. Some endpoints, such as volume of distribution, have been shown to be modelable across multiple studies, while for other endpoints, such as clearance and in particular bioavailability, results differ more widely, and they are generally less satisfactory. However, what is common to the above studies is that models were generally either based on limited compound data sets and/or the number of PK endpoints modeled was limited to a small number.

In this work, we describe machine learning models that predict several rat in vivo PK parameters: F, CL, volume of distribution at the steady state (V_{ss}), AUC, C_{max}, and $t_{1/2}$, and concentration−time PK curves. The models are trained and validated on a large data set of more than 3000 compounds. The combination of endpoints modeled and the quantity of in vivo data used for training, to the best of the knowledge of the authors, make it the most comprehensive model of its type, both in the output property space and with respect to chemical space coverage. Given the nonclinical nature of the data set, the endpoint value ranges are wider than those for successful drugs, ensuring better model coverage across the value range. We explore state-of the-art AI approaches, such as graph convolutional neural networks that encode the molecular chemical graph structure,[26] as well as traditional machine learning algorithms utilizing molecular property descriptors. In addition to chemical descriptors, the models use several in vitro ADME properties as input features. Various imputation approaches for missing in vitro data, including utilizing corresponding in silico predictions or using deep learning technology[27] able to handle sparse and noisy experimental data, were explored. In addition, we will describe a deep neural network model for concentration−time PK profiles and compare the accuracy of PK parameters estimated from predicted PK curves versus the accuracy of PK parameter model predictions.

2. METHODS

2.1. Data Set. In vivo rat PK data (iv and po administration) were extracted from the internal AstraZeneca database. To ensure data consistency, only data generated in male Han Wistar rats since 2013, at a single investigation site, were used. The data set focused on low-dose PK studies, that is, the majority of compounds (>92%) were dosed <5 μmol/kg iv and <10 μmol/kg po. At least two replicates (i.e., two animals) for each administration route were available per compound. Nine PK parameters were extracted for modeling; five parameters corresponded to the iv route: AUC iv, C_{max} iv, $t_{1/2}$ iv, CL, and V_{ss}, and four parameters corresponded to the po route: AUC po, C_{max} po, $t_{1/2}$ po, and F, defined as the percentage of a po dose that reaches the systemic circulation, given by the following equation

$$F(\%) = \left(\frac{\mathrm{AUC_{po}}}{\mathrm{AUC_{iv}}} \cdot \frac{D_{iv}}{D_{po}}\right) \cdot 100$$

where D_{iv} and D_{po} are iv and po administered doses, respectively. C_{max} for an iv experiment corresponds to C_0.

In addition, dose-dependent time−concentration curves were extracted from the iv and po routes, spanning a time period of 2 min to 24 h.

2.1.1. In Vivo Experimental Details. Male Han Wistar rats, aged 6−8 weeks, were dosed either via the tail vein (iv) or oral gavage (po). Compounds were dosed in cassettes of up to five compounds at low doses (see above). Standard formulations for iv administration were solutions containing cyclodextrin or other solubilizing agents in acceptable quantities, whereas for po administrations, suspensions using hydroxypropyl methylcellulose (HPMC) were usually preferred. Blood samples were taken at predefined timepoints after dosing, usually 10 occasions up to 24 h, collected in ethylenediaminetetraacetic acid (EDTA)-containing tubes, and centrifuged at 4000g for 5 min at 4 °C to obtain plasma. Plasma samples were stored at −75 °C until they were analyzed using liquid chromatography−tandem mass spectrometry (LC−MS/MS). The resulting time−concentration profiles were evaluated using noncompartmental analysis (NCA).

2.1.2. Data Curation. The AUC (μM*h), C_{max} (μM), and concentration values (μM) were scaled by the dose (μmol/kg). Two formats of the data were considered—aggregated (where the values of the PK parameter were averaged between replicates) and non-aggregated (where each compound had several replicate values for the PK parameter). The time−concentration curves were non-aggregated (majority of compounds had two curves per each administration route). Compounds with the molecular weight higher than 750 Da were excluded from the data set. The final data set consisted of 3070 compounds.

2.1.3. Data Transformations. AUC iv and po [μM*h/(μmol/kg)], C_{max} iv and po [μM/(μmol/kg)], CL [mL/min/kg], and V_{ss} [L/kg] were $\log_{10}$-transformed. To be able to include zero values in the analysis, a minimum cutoff value a_{min} was defined in the log-transformed space for each of these parameters (based on the data spread): $a_{min} = -4$ for AUC (iv and po) and C_{max} iv, $a_{min} = -5$ for C_{max} po, $a_{min} = -0.5$ for CL, and $a_{min} = -2$ for V_{ss}. No transformation was applied to half-life (h) iv and po. F was first normalized by the maximum value in the data set (F = 160%), normalized values below 0.01 were set to 0.01, and then, the logit transformation was used, where logit $y = \log_{10}(y/(1-y))$. Concentration values in time−concentration profiles were $\log_{10}$-transformed, and no a_{min} cutoff was applied. The distributions of transformed PK parameters are provided in Figure S1 of the Supporting Information.

2.1.4. Experimental Variability of the Measurements. The experimental variability present in the data was estimated by calculating the standard deviation between replicate measurements for each compound with more than two replicates and taking the 95%-quantile of the distribution of standard deviations as the estimate for the experimental noise/error. 95%-quantile was selected arbitrarily with the rationale to capture majority of values for standard deviations without considering more extreme values.

For the bioavailability parameter, it is not expected that the bioavailability would be the same at high doses as that at low doses. Since the data set focused on low-dose PK studies (92% of compounds were dosed <5 μmol/kg iv and <10 μmol/kg po), we did not anticipate differences in bioavailability values due to differences in doses, and the difference in values was attributed to variability in subject/animal responses. Therefore, all replicates per compound for all doses were treated equally in the estimation of experimental variability.

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

2.1.5. In Vitro ADME Properties. Nine experimentally obtained ADME and physicochemical properties were added to the data set to be used as input features to the model. These in vitro data points were collected prior to the in vivo studies and are often implemented early in lead optimization. The properties describe compound lipophilicity, solubility, permeability, and active transport properties; intrinsic metabolic clearance; and plasma protein and hepatocyte binding:

- LogD
- Solubility [dried dimethyl sulfoxide (DMSO)]
- Caco-2 intrinsic permeability
- Caco-2 efflux ratio
- Human liver microsome intrinsic clearance
- Rat hepatocyte intrinsic clearance
- Rat plasma protein binding
- Human plasma protein binding
- Fraction unbound in rat hepatocytes

Log-transformed values were used for Caco-2 intrinsic permeability, Caco-2 efflux ratio, human liver microsome, and rat hepatocyte intrinsic clearance values. Rat and human plasma protein binding, as well as the fraction unbound in rat hepatocytes, were logit-transformed. If multiple measurements existed for a compound, the replicate values were averaged by using the arithmetic mean for log-transformed properties (post-transformation) and the median for binding values. Overall, about 25% of the in vitro values were missing in the data set. The assay-dependent percentage of missing values ranged from 6% (LogD) to 55% (fraction unbound in rat hepatocytes). The distributions of the in vitro input features are provided in Figure S2 of the Supporting Information.

Both human and rat in vitro clearance and plasma protein binding were included as input features. This is because both human and rat endpoints are relevant to PK concepts. Another reason for the inclusion of human in vitro properties was a different level of missing data in various properties. For example, 19% of rat plasma protein binding values were missing versus 11% of human plasma protein binding values. By including human in vitro properties, fuller coverage was ensured.

2.1.6. In Vitro ADME Experimental Details. In vitro properties were measured in routine high-throughput assays: LogD was measured using a shake flask method in 96-well plates.[28,29] Solubility was measured as thermodynamic solubility from DMSO stock solution, where DMSO was evaporated before analysis, again using a shake-flask method.[28,30] Caco-2 intrinsic permeability was measured in the presence of a transporter inhibitor cocktail, considering a pH gradient of a pH of 6.5 at the apical side and a pH of 7.4 at the basolateral side, whereas pH was 7.4 on both sides with no inhibitor cocktail when measuring the Caco-2 efflux ratio.[31] Intrinsic clearance was determined in high-throughput assays using incubations of cryopreserved human microsomes or rat hepatocytes at 37 °C for up to 60 or 120 min, respectively.[32−34] Plasma protein binding data were generated using equilibrium dialysis.[33,35,36] The fraction unbound in rat hepatocytes was also determined using equilibrium dialysis.[37]

2.1.7. In Silico Predictions of In Vitro ADME Properties. Predictions for the ADME and physicochemical properties listed in Section 2.1.5 were added to the data set. The models for these properties were developed using large internal data sets (≥4000 compounds in smaller data sets and up to 160,000 compounds in the larger data sets). Models for the Caco-2 intrinsic permeability and Caco-2 efflux ratio were developed using the random forest algorithm with OESelma molecular property descriptors[38] (see Section 2.2.1). Scikit-learn implementation was used for random forest.[39] The rest of the properties were modeled using a support vector machine with signature descriptors[40] and the conformal prediction framework[41] implemented in the CPSign software.[42,43] A temporal test set (10% of the data) was used for validation, where a data set was split chronologically into training and test sets and 10% of the latest data were reserved for the test set. The approach represents a real-life scenario of model usage. The models are regularly updated, with the frequency of update varying between 1 and 6 months, depending on the amount of data being generated for each property. Model performance is monitored continuously by predicting the new data before each model update. Details of model performance and methods were described recently by Oprisiu and Winiwarter.[44]

2.1.8. Missing Data Imputation. As mentioned in Section 2.1.5, around 25% of the in vitro ADME property values were missing. Since the majority of machine learning algorithms require all feature values to be present, two approaches for the imputation of missing values were adopted. The first approach, further on referred to as the "replace" approach, was to replace missing in vitro values with corresponding in silico predictions. The second approach was an imputation approach built-in within the Alchemite method,[27,45] referred to as the "impute" approach, as described below in Section 2.3.5.

2.1.9. Training/Test Data Set Split. A temporal split was used to divide the data into the training and test sets, that is, around 10% of compounds (312 compounds) with the latest synthesis date were separated into the test set. The test set was not used during training and hyperparameter optimization. Table 1 describes number of compounds in the training and test sets for all endpoints. The multitask format here refers to an approach where a single model is built for all nine PK parameters.

Table 1. Number of Compounds/Rows in the Training and Test Sets for the Aggregated and Non-aggregated Data Formats

endpoint	N train	N test
Aggregated Format		
AUC iv	2686	312
AUC po	1822	261
F	1817	266
CL	2682	312
C_{max} iv	2689	312
C_{max} po	1899	273
$t_{1/2}$ iv	2685	312
$t_{1/2}$ po	1755	256
V_{ss}	2686	312
overall (multitask format)	2758	312
Non-aggregated Format		
concentration of dose−time profile iv	5895	632
concentration of dose−time profile po	4266	578

The diversity of the chemical space of the training and test sets was analyzed by considering Bemis−Murcko scaffolds. Overall, the set contains 1845 scaffolds, with 1644 scaffolds

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

unique to the training set, 173 scaffolds unique to the test set, and 28 scaffolds common between training and test sets. The 28 scaffolds represent less than 2% of the training set scaffolds and correspond to 77 compounds (25%) of the test set. Therefore, 75% of compounds in the test set have scaffolds different from the training set scaffolds, indicating a low overlap between the two sets.

2.2. Chemical Descriptors. *2.2.1. OESelma Molecular Properties.* The OESelma descriptors were generated by AstraZeneca's in-house program OESelma.[38] They comprise around 100 common 1D and 2D molecular descriptors related to physicochemical properties, such as size, ring structure, flexibility, atom types, hydrogen bonds, polarity, electronic environment, partial atom charge, and lipophilicity, including connectivity indices.[46] Additionally, LogD and LogP (base-10 logarithm of partition coefficient) from ACDLabs[47] and LogP from Biobyte[48] were included in the descriptor set. These descriptors have been shown useful in QSAR modeling, see, for example, studies by Bruneau,[49] Wood et al.,[50] and Fredlund et al.[31]

2.2.2. Chemprop Graph Convolutions. In contrast to traditional chemical descriptors, graph convolutional neural networks learn how to represent molecules directly from chemical structure in an end-to-end learning fashion.[51,52] In this study, the directed message passage neural network framework (D-MPNN) Chemprop[26] was used. Chemprop consists of a message-passing phase that creates molecular representations using a graph convolutional neural network and a readout phase that learns and predicts the final endpoints. The D-MPNN is initialized with a set of atom features (atom type, number of bonds, formal charge, chirality, number of bonded hydrogen atoms, hybridization, aromaticity, and atomic mass) as nodes and bond features (bond type, conjugation, ring membership, and geometric isomerism) as edges in a graph representation. From the graph, messages are created from the bond vectors, which continuously update the molecular representation based on the neighboring atom vectors. The weights and biases for this network are updated during training, and the hyperparameters are optimized as described in Section 2.3.1 covering the readout phase.[26]

2.2.3. Signature Descriptors. Molecular signatures[40] are 2D descriptors, which combine all atomic signatures of a molecule. An atomic signature is a canonical representation of the atom's environment up to a predefined connectivity, denoted as height. Signature CPSign implementation was used[42] with default settings. Signature heights were ranged from 0 to 3.

2.2.4. StarDrop Descriptors. The descriptors were calculated with the Auto-Modeller module of StarDrop software[53] using SMILES strings defining the structure of each compound. A total of 330 descriptors were calculated, including whole-molecule properties such as molecular weight, LogP, and polar surface area; and 2D structural fragments were defined by SMARTS strings.[54]

2.3. Description of Modeling Techniques. *2.3.1. Chemprop.* The readout phase of Chemprop is a feed-forward neural network.[26] Five-fold cross-validation based on scaffold splits was performed using the built-in hyperparameter optimization functionality to optimize a set of hyperparameters: size of the layers in the convolutional neural network, number of message-passing steps, dropout and number of layers in the feed-forward networks. The scaffold splitting ensures that each molecular scaffold, calculated using the RDKit implementation of Bemis−Murcko decomposition, only appears in one of the splits.[55] As a result, the cross-validation performance is based on an unseen chemical space, which is similar to how models are used in an industrial setting. ReLU (rectified linear unit) was chosen as the activation function.[56] Five models with the same architecture but different parameter initializations were trained for 70 epochs and used as an assembly, providing uncertainty in prediction as well as prediction values. The average of predictions of individual ensemble models was taken as the predicted value, and the standard deviation between individual predictions estimated the uncertainty. The algorithm was used to build both single-task and multitask models, where nine PK parameters represented multiple tasks. In addition to the graph convolutions, in vitro ADME properties with missing values replaced with corresponding in silico predictions ("replace" approach) were added to the final feature set.

2.3.2. Gaussian Processes Regression. Gaussian Processes is a kernel-based Bayesian probabilistic method,[57,58] which was previously successfully utilized for ADME and PK modeling.[23,24,59−61] MATLAB 2019a implementation was used in this work.[62] Five kernel functions were explored: exponential, squared exponential, rational quadratic, automatic relevance determination (ARD) squared exponential, and ARD exponential. For the rest of the hyperparameters, the defaults were accepted. 10-fold random-based cross-validation was used to supervise model performance. The algorithm was used with OESelma descriptors and in vitro ADME properties ("replace" approach for missing values).

2.3.3. Gradient Boosting Regression. Gradient tree boosting is an algorithm that produces an ensemble of weak decision trees and can be used both for regression and classification. It is a generalization of adaptive boosting to arbitrary differentiable loss functions. The boosting works in an additive way, where weak learners are added one at a time and the optimization is driven by a gradient descent-like procedure. Gradient boosting regression as implemented within Scikit-learn was used.[39] Grid search with five-fold random-based cross-validation was used to optimize hyperparameters and to supervise model performance in training. The algorithm was used with OESelma descriptors and in vitro ADME properties ("replace" approach for missing values).

2.3.4. Support Vector Machine—CPSign. CPSign algorithm[42] is a support vector machine (SVM) with signature descriptors[40] and a conformal prediction framework.[41] The radial basis function (RBF) kernel was used in the models with default values for hyperparameters. Five-fold random-based cross-validation was used to supervise model performance and to perform calibration.

2.3.5. Alchemite. Alchemite is an imputation and prediction method designed to handle sparse input data that have been used in a variety of chemistry and materials science domains.[27,45,63] It is a deep neural network method. In this work, it was used to predict either PK parameters, in common with the methods described above, or PK curves directly. In both cases, Alchemite used an ensemble of 200 sub-learners trained on random subsets of the available training data, with the resulting prediction being the average of the ensemble's predictions and the sub-learners' variance giving an estimate of the uncertainty. Alchemite was run for predicting PK parameters in the multitask format using three different

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

classes of input data: the "ivo" approach used only in vitro only data as the input, which were sparse and so was imputed as part of the model training; the "ivis" approach used both sparse in vitro data and complete in silico data as the input, relying on Alchemite to identify the correlations between the data sets to impute the gaps in the in vitro data; and the "replace" method, where the missing in vitro values were directly filled using in silico results (see Table 2). In all cases, five-fold random-split cross-validation was used to optimize hyperparameters using the Bayesian tree of Parzen estimators algorithm.[64]

Alchemite was used to build models directly of PK concentration−time curves as well as PK parameters. Both iv and po dosing PK curves were modeled simultaneously, using the "replace" approach to deal with the missing in vitro data. Alchemite uses the measurement time as an additional input when modeling curve data, creating a list of time points for each curve and associating these with an equal-length list of concentrations, in parallel for the iv and po curves. During the training time, these lists are expanded into multiple training data points on the fly, ensuring that curves with different numbers of data points are weighted equally by the algorithm (to avoid putting more emphasis on curves with more measurement points). At the prediction time, an arbitrary list of time points can be evaluated in parallel.

In the experimental concentration−time data, many points were missing as the measured concentration fell below the measurement tolerance: for modeling purposes, these points were replaced by the minimum measured concentration in the data set (4.3×10^{-6} μM/μmol/kg) to ensure that the model was aware of the tendency to low concentrations at late times. The log-concentration was modeled to provide accuracy over multiple orders of magnitude of concentration.

2.3.6. Combinations of Algorithms and Descriptors. Not all combinations of descriptors and modeling techniques were investigated. Table 2 describes the approaches and algorithms that were explored for PK parameter modeling and specifies abbreviations used for various techniques. Neural network (NN) methods, Chemprop and Alchemite, were used in a multitask format. Chemprop was also used in a single-task format, and this format was also utilized by the rest of the algorithms. Only the Alchemite algorithm was used to model concentration−time PK profiles (for details, see Section 2.3.5).

2.4. Evaluation of Uncertainty Estimates. Two metrics were considered to evaluate the quality of different uncertainty estimates—ranking-based and calibration-based.[65]

2.4.1. Ranking-Based Confidence Curve. To construct the confidence curve, the compounds were ordered by the predicted uncertainty in a decreasing order. The compounds with the highest uncertainty are gradually removed and the RMSE (root mean squared error) is measured for the remaining subset. The RMSE of the subset $[(100 - n)\%$ of compounds with the lowest uncertainty$]$ is plotted as a function of the confidence percentile n.[65] The so-called "oracle" confidence curve represents a perfect situation, where the true error is used to order the compounds. In the ideal scenario, the confidence curve is as close as possible to the oracle curve, which represents a lower bound. The area under the confidence−oracle error, AUCO, which is defined as the difference between the areas under the both curves, can be used as a quality metric.

Table 2. Combinations of Features, Algorithms, and Approaches Explored for PK Parameter Modeling Together with the Respective Abbreviations

	algorithm	Chemprop multitask (MT)	Chemprop single-task (ST)	gradient boosting regression	gaussian processes	support vector machine	Alchemite multitask
	input features	graph convolutions ADME properties	graph convolutions ADME properties	OESelma descriptors ADME properties	OESelma descriptors ADME properties	signature descriptors OESelma descriptors ADME properties	StarDrop descriptors ADME properties
				Aggregated Format			
use of ADME features and missing value approaches	ADME in vitro ("replace" for missing values)	Chemprop MT	Chemprop ST	GBoost	GPR	CPSign	Alchemite (replace)
	ADME in vitro only ("impute" for missing values)						Alchemite (ivo)
	ADME in vitro ("impute" for missing values) + ADME in silico						Alchemite (ivis)
				Non-aggregated Format			
use of ADME features and missing value approaches	ADME in vitro ("replace" for missing values)						Alchemite (replace) nAgg
	ADME in vitro only ("impute" for missing values)						Alchemite (ivo) nAgg
	ADME in vitro ("impute" for missing values) + ADME in silico						Alchemite (ivis) nAgg

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

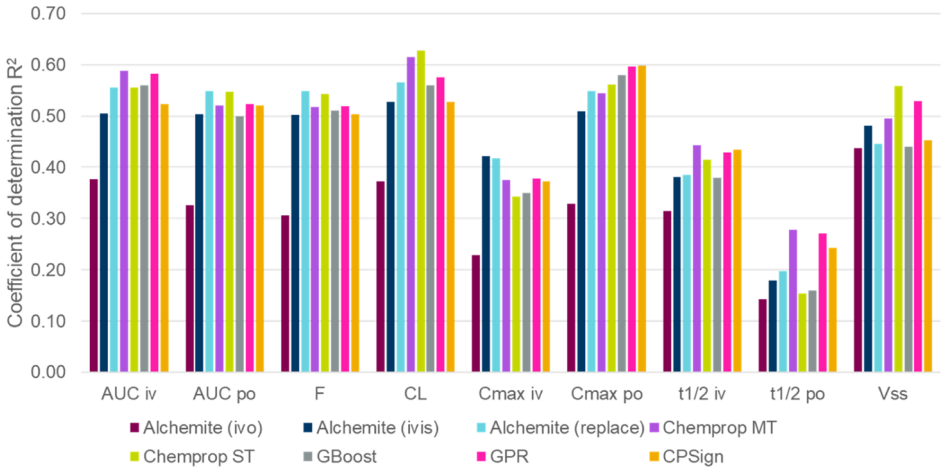


Figure 1. Coefficient of determination (R^2) on the test set for the nine PK parameters using different models built using the aggregated data format. Alchemite (ivo) is an Alchemite multitask NN algorithm with in vitro features only and imputation (mulberry bars) and Alchemite (ivis) is an Alchemite algorithm with in silico and in vitro features and Alchemite imputation of missing in vitro values (dark blue bars). The rest of the techniques use in vitro features, where missing values are replaced with in silico values ("replace" approach). Alchemite (replace) is an Alchemite algorithm (light blue bars), and Chemprop MT and Chemprop ST are Chemprop NNs in multitask and single-task modes (purple and green bars, respectively), GBoost is a gradient boosting regression (gray bars), GPR is a Gaussian Processes regression (pink bars), and CPSign is an SVM conformal regression (orange bars).

2.4.2. Calibration Curve. In the calibration curve, the actual values of predicted uncertainty are used as opposed to the ranking order only. In interval-based calibration, it is assumed that each prediction and its uncertainty correspond to the mean and the standard deviation of a Gaussian distribution defining predictive distribution. To build a calibration curve, the confidence value is varied between 0 and 1. For each confidence value, the symmetric confidence interval around the mean is defined (for a fixed confidence, the interval around the mean would be different for each compound because the standard deviation defined by uncertainty is compound-dependent). Then, it is calculated for how many compounds the observed values fall in the corresponding confidence interval of the predictive distribution, that is, the empirical probabilities of belonging to each interval. In a perfectly calibrated model, n % of the predictions would fall in the n-th confidence interval, resulting is a diagonal line for a perfect calibration curve. In a well-calibrated model, the calibration curve is close to the diagonal line. The area under the calibration error curve, AUCE, which is defined as the absolute difference between the areas under the calibration and perfect curves, can be used as a quality metric.[65]

Two calibration curves, corresponding to two values of uncertainty, were considered. In one case, the uncertainty predicted by the model σ_m was used to construct the curve. In the second case, the uncertainty due to variability in experimental measurements, also called aleatoric uncertainty, was added to the model uncertainty to define the total uncertainty σ_{total} as follows

$$\sigma_{total}^2 = \sigma_m^2 + \sigma_{exp}^2$$

where σ_{exp} is the experimental error.

2.5. Description of the WSM. Hepatic elimination remains the primary route of elimination for drugs;[66] hence, IVIVE using the WSM is routinely applied.[13,15,67,68] The WSM is a mathematical model of the liver and requires intrinsic clearance from hepatocytes or liver microsomes as input parameters. If CL prediction accuracy is high and a mechanistic understanding of compound CL in animals can be achieved, this provides a level of confidence for extrapolation to humans. Hepatic metabolic clearance is calculated as follows using the WSM

$$\text{clearance(hepatic metabolic)(mL/min /kg)} = \frac{\left(\text{CL}_{\text{int},u} \times \left(\frac{fu_p}{\text{Rb}}\right) \times \text{Q}_h\right)}{\left(\text{CL}_{\text{int},u} \times \left(\frac{fu_p}{\text{Rb}}\right) + \text{Q}_h\right)}$$

where $\text{CL}_{\text{int},u}$ is the scaled unbound intrinsic clearance; fu_p is the fraction unbound in plasma; Rb is the blood/plasma ratio, and Q_h is the hepatic blood flow.

2.6. Calculation of PK Parameters from Predicted Concentration−Time Profiles. PK parameters were calculated from predicted concentration−time curves via NCA using SimBiology App of MATLAB R2019a.[62,69] Predicted values that fell below half the minimum of the experimentally observed values (4.3×10^{-6} μM/μmol/kg) were removed to aim for consistency with the experimental results in the treatment of low concentrations.

3. RESULTS AND DISCUSSION

3.1. PK Parameter Models. *3.1.1. Summary of Results.* The purpose of this work was to build an accurate and useful model of the PK parameters and not to compare different machine learning algorithms, descriptors, and approaches to each other. Therefore, only selected combinations of descriptors and modeling techniques were investigated (described in Section 2.3.6 and Table 2).

The results of modeling efforts for the aggregated data format are summarized in Figure 1 which shows the coefficient of determination (R^2) evaluated on the test set. The detailed results including RMSE on the test set are shown in Figure S3 and Table S1. Models with an acceptable accuracy ($R^2 > 0.5$) were achieved for the majority of the endpoints, except for C_{max} iv, $t_{1/2}$ iv, and $t_{1/2}$ po. Figure 1

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

Table 3. Best Model for Each PK Parameter Together with the Coefficient of Determination (R^2), RMSE, Fold-Error (in the Nontransformed Space), and Percentage of Compounds with the Error within 2- and 3-Fold in the Nontransformed Space

PK parameter	best model (s)	R^2	RMSE	fold-error	% within 2-fold error	% within 3-fold error
AUC iv	Chemprop MT = GPR	0.59	0.28	1.9	76	93
AUC po	Alchemite (replace), Chemprop ST	0.55	0.61	4.1	54	68
F	Alchemite (replace), Chemprop ST	0.55	0.46	2.9	65	84
CL	Chemprop ST, Chemprop MT	0.63	0.26	1.8	78	94
C_{max} iv	Alchemite (ivis), Alchemite (replace)	0.42	0.22	1.7	87	97
C_{max} po	GPR = CPSign	0.60	0.56	3.7	44	60
$t_{1/2}$ iv	Chemprop MT	0.44	1.84		55	78
$t_{1/2}$ po	Chemprop MT	0.28	2.30		80	95
V_{ss}	Chemprop ST	0.56	0.27	1.8	78	93

shows that different techniques result in models of similar accuracy and there is no single technique that exceeds other methods across all endpoints. Alchemite (ivo) and Alchemite (ivis) use the Alchemite method of imputation, based, respectively, on in vitro data only and in vitro data supplemented by in silico data. The rest of the models use a "replace" approach, where missing in vitro values are replaced with in silico values. The results (Figure 1) show that the models using the "replace" approach generally outperform models using imputation. Performance of the Alchemite (ivis) model, which uses the built-in Alchemite imputation method to impute missing in vitro parameters and also includes in silico features, closely follows the performance of the "replace" models; the R^2 values are slightly lower than those of the corresponding "replace" models, except for C_{max} iv. For this endpoint, the Alchemite (ivis) model showed the highest R^2 value of all methods (R^2 = 0.42), even though the difference from the Alchemite (replace) method was minor, and C_{max} iv was one of the endpoints with overall less accurate models. It is hard to know how much the "imputed" in vitro features are used in the models since the highly correlated in silico features are available in the descriptor set. (Building the model using only in silico features showed equivalent performance. Data are not shown here.) Alchemite (ivo) represents the imputation of in vitro ADME values in the absence of ADME in silico predictions and tests the power of a "true" imputation approach in a scenario where predictive models of in vitro properties are not available. It underperforms in comparison with models using the "replace" approach. This suggests that the in silico models trained on a large set of ADME data are more accurate than relying on imputation within a smaller project data set, which aligns with our expectations.

Focusing on the models using the "replace" approach, for the majority of endpoints, neural network algorithms Alchemite (replace), Chemprop MT, and Chemprop ST yield the best performing model with the exception of C_{max} po, where the Gaussian Processes model (GPR) provides the best performance with R^2 = 0.6 (see Figure 1 and Tables S1 and 3). The single-task neural network models provide broadly equivalent performance to multitask models on most of the endpoints; for V_{ss}, Chemprop ST performed better than others (R^2 = 0.56). The traditional machine learning algorithms, Gaussian Processes and gradient boosting, closely follow neural network models in performance for most of endpoints. The SVM with a conformal regression technique (CPSign) underperforms for many endpoints. A possible explanation is that the automatic model building procedure used in CPSign is designed for the signature descriptors and—without adaptation—is not so well-suited for other descriptor types such as in vitro ADME properties. It should be noted that there is a slight variability in performance of models built by different runs for all techniques apart from the GPR due to a different initialization of weights in neural network methods and different (random) cross-validation splits, which would in turn affect hyperparameter optimization. Due to this variability, which was not fully captured, the performance of all "replace" algorithms apart from the CPSign can be considered equivalent.

The best model for each endpoint was selected based on the lowest RMSE (selection on the highest R^2 produces the same results) on the test set, and the results are shown in Table 3. The models where the difference between the RMSE and the lowest RMSE did not exceed 0.005 were considered of similar performance. While RMSE represents an error in the log-transformed space, a corresponding fold-error provides an estimation of error in the nontransformed space. The fold-error is not applicable to half-life models since $t_{1/2}$ iv and po were not log-transformed. For bioavailability, the fold-error is an upper-bound estimate since logit transformation was used instead of log.

The Alchemite method was also applied to the non-aggregated data set, where each compound had several replicate values of the PK parameter. The results are shown in Figure S4. The use of the non-aggregated data does not present any advantages. For the majority of the endpoints, the performance of models based on that format is slightly lower than or equivalent to the performance of models based on the aggregated format.

Since bioavailability and clearance represent the most important PK parameters for decision making in projects, the models for these are explored in more detail in the following subsections.

3.1.2. Bioavailability Model. The best model for bioavailability was produced using the Alchemite (replace) method, a multitask deep neural network with 2D chemical descriptors, where missing in vitro features were replaced with in silico values. The Chemprop single-task model (Chemprop ST) produced equivalent results (see Tables 3 and S1). The model achieved a good performance on the temporal test set of 312 compounds, with R^2 = 0.55 and RMSE = 0.46. The experimental error is estimated at 0.43 (in the logit-transformed space). The RMSE of the model is close to the level of experimental error. The achieved RMSE corresponds to a roughly 2.9 fold-error in the nontransformed space and to RMSE = 0.37 in the log-transformed space. The scatter plot of predicted versus observed values for logit-transformed F is shown in Figure 2. 65 and 84% of compounds are predicted

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

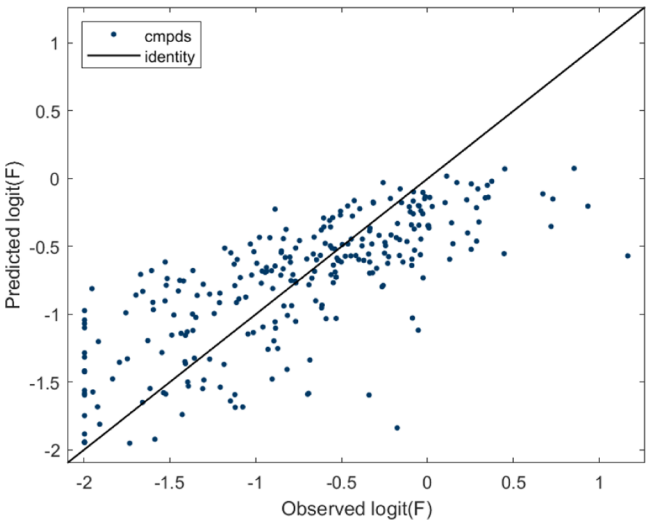


Figure 2. Predicted vs observed values for logit(F) on the test set for predictions made using the Alchemite (replace) model. The identity line is a solid black line.

within 2- and 3-fold from the experimental value (in the nontransformed space), respectively. The performance of the model makes it well-suited to help decision making in early drug discovery. To compare to the published results, Schneckener et al.[20] reported that a model for oral bioavailability in rats achieved R^2 = 0.18 and RMSE = 1.04 (in the log-transformed space); the model, based on ∼1900 compounds, utilized a deep neural network approach and a chemical structure as the input (converted to descriptors using a prebuilt neural network).

3.1.3. Clearance Model. Clearance is one of the most challenging parameters to optimize in drug discovery. Low clearance is desired for a drug candidate to achieve acceptable duration of target engagement. The best model for CL was produced using the graph convolutions neural network method Chemprop applied in a single-task setting (Chemprop ST), with the Chemprop multitask model (Chemprop MT) producing equivalent results (see Tables 3 and S1). The model achieved a good performance on the temporal test set of 312 compounds, with R^2 = 0.63 and RMSE = 0.26. The RMSE of the model is only slightly higher than the experimental error estimated at 0.18 (in the log-transformed space) and corresponds to a 1.8 fold-error in the non-transformed space. The scatter plot of predicted versus observed values for log-transformed CL is shown in Figure 3A. 78 and 94% of the compounds are predicted within 2- and 3-fold error (in the nontransformed space), respectively.

This performance compares favorably to published results. Feinberg et al.[21] reported Pearson's r^2 = 0.275 for a rat CL model built using a single-task graph convolution neural network technique on a data set of ∼60,000 compounds utilizing a temporal-split test set. A rat CL model based on 1114 compounds utilizing an RBF technique and several in vitro ADME properties in addition to 2D descriptors achieved R^2 = 0.61 and RMSE = 0.31 on a cluster-split test set;[23] it is anticipated that the performance would be slightly lower on the temporal-split test set.

3.1.4. Comparison with the WSM. The WSM is a standard IVIVE tool for estimation of hepatic clearance. It is routinely applied in decision making for compound prioritization and progression for in vivo testing and also to gain an understanding of the mechanism of clearance.[13,67,68] In drug discovery, and in the absence of other data, results from the WSM are used as an approximation of the total clearance. Figure 3 shows a comparison of predictions from the CL model and WSM results on the test set of 312 compounds. Predictions of the WSM are restricted by the rat liver blood flow [Q_h = 72 mL/min/kg[32] or $\log_{10}(Q_h)$ = 1.86]; therefore, the WSM predictions were available only for 259 compounds of the test set. As seen from Figure 3B, the WSM model significantly underpredicted the total clearance on this set, achieving R^2 = −0.11 and RMSE = 0.44. The squared Pearson's correlation coefficient, r^2, between predicted and observed values is 0.51, showing that the correlation is high but the magnitude of the predicted values is underestimated. The CL model provided much better accuracy with R^2 = 0.63 and RMSE = 0.26, (r^2 = 0.63). Therefore, the CL model provides an accurate and useful tool for decision making in

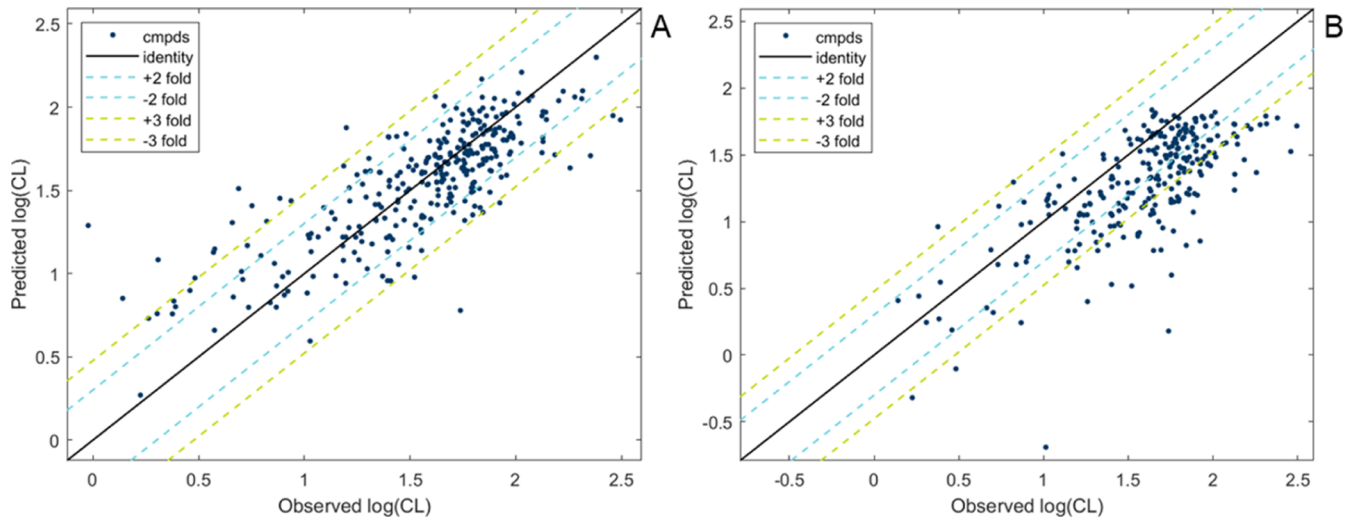


Figure 3. Predicted vs observed values for log(CL) on the test set for (A) predictions made using the Chemprop ST model and (B) predictions made using the WSM (well-stirred model) (259 compound subset of the test set). The identity line is a solid black line, $\pm\log_{10}(2)$ lines corresponding to a 2-fold error are dashed blue lines, and $\pm\log_{10}(3)$ lines corresponding to a 3-fold error are dashed green lines.

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

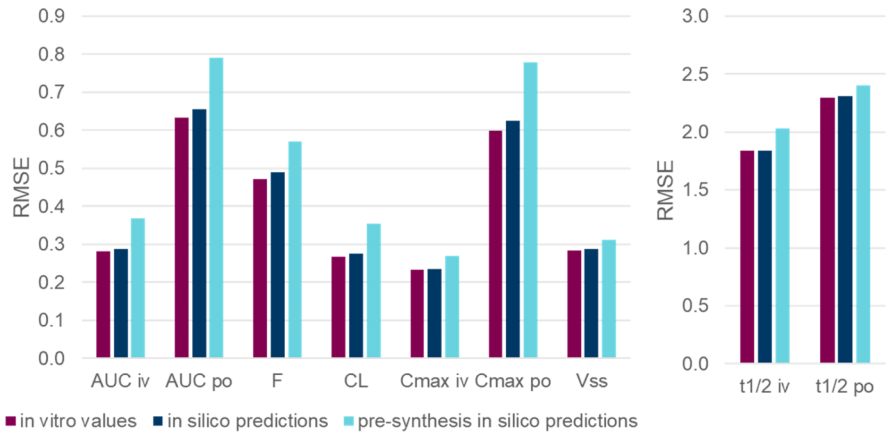


Figure 4. Performance of the Chemprop MT model on the test set utilizing in vitro measurements for ADME features or corresponding in silico predictions. RMSE on the test set is shown when in vitro values (mulberry bars), in silico predictions (navy bars), and pre-synthesis in silico predictions (light blue bars) are used for ADME and physicochemical features.

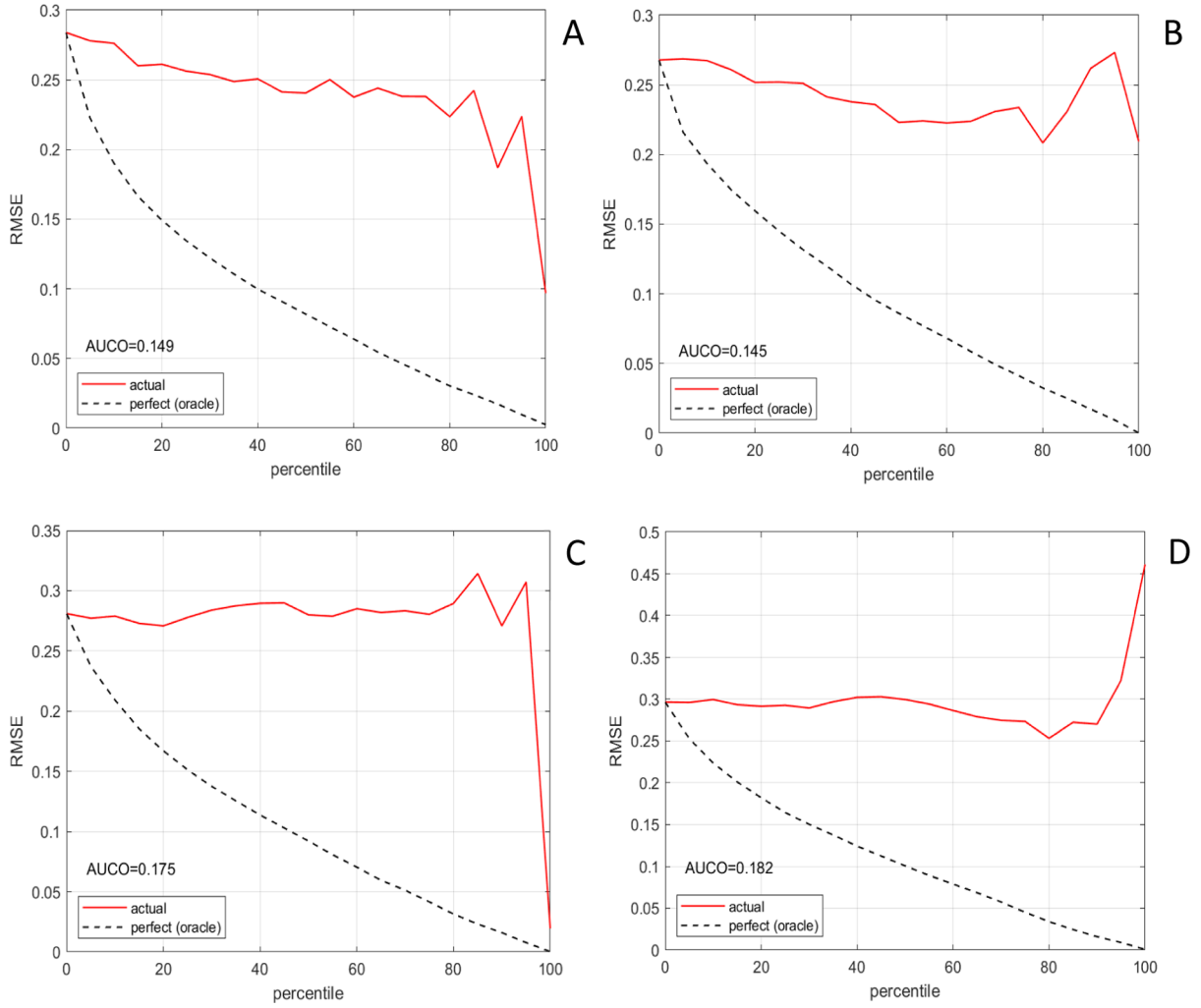


Figure 5. Confidence curves and corresponding AUCO values for the CL endpoint obtained on the test set using predictions and uncertainty estimates from different models—(A) Alchemite (replace), (B) Chemprop MT, (C) GPR, and (D) CPSign. The oracle curve is a dashed black line.

early discovery to guide compound prioritization and selection. Also, the CL model is not restricted by the liver blood flow and can predict compounds with high clearance. Its application is complementary to the WSM, and the agreement or disagreement of predictions from both models can inform on the mechanism of clearance.

3.1.5. Predicting Compound PK at the Point of Design. In order to test whether the models can be used at the point of

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

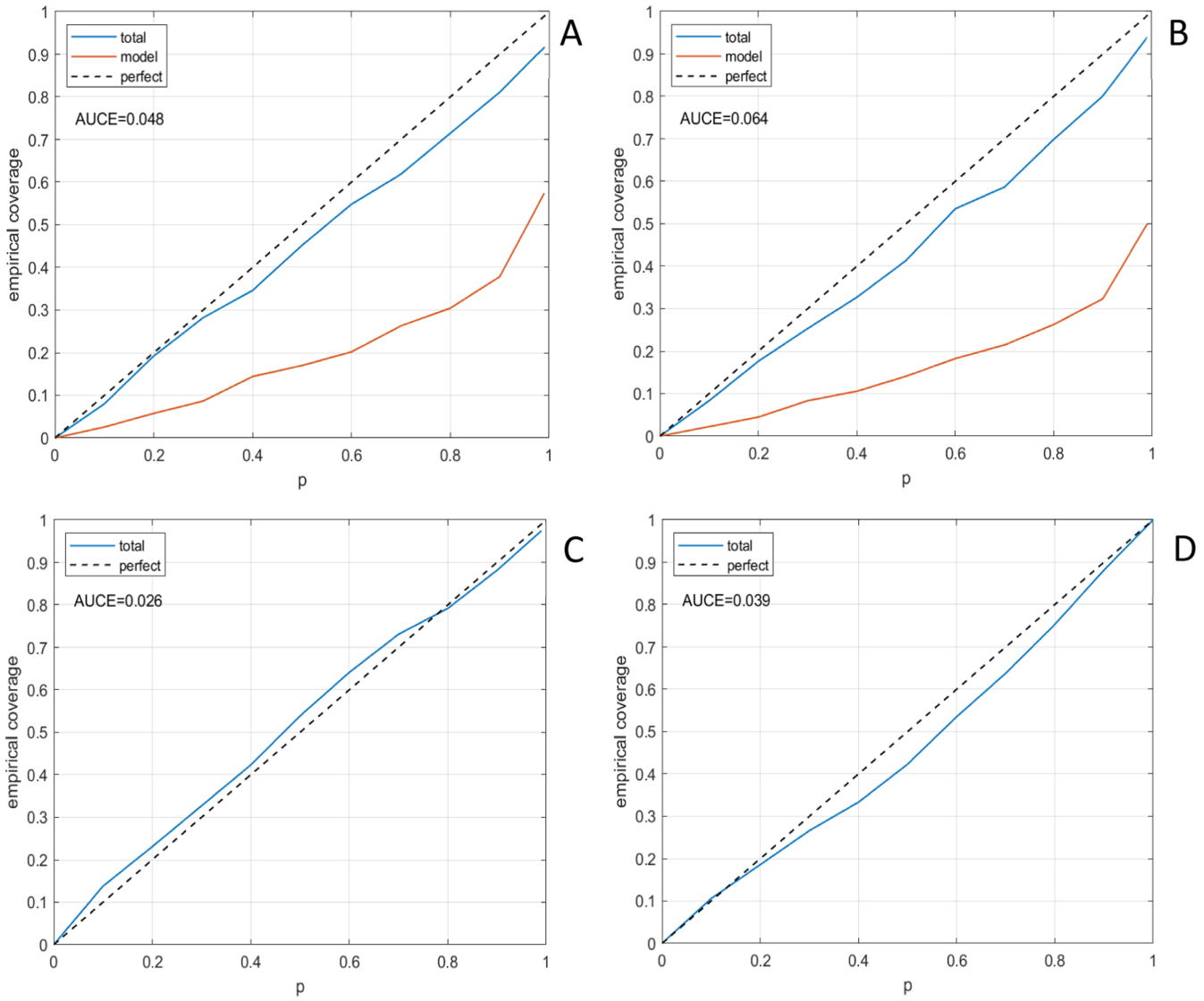


Figure 6. Calibration curves and corresponding AUCE values for the CL endpoint obtained on the test set using predictions and uncertainty estimates from different models—(A) Alchemite (replace), (B) Chemprop MT, (C) GPR, and (D) CPSign. The confidence curves based on the model uncertainty and the total uncertainty are red and blue lines, respectively. The perfect calibration curve is a dashed black line. The AUCE value corresponds to the total uncertainty.

design, before compounds are synthesized and when ADME in vitro properties are not available, the performance of the Chemprop MT model was evaluated on the test set in the following two scenarios. First, in silico predictions were used instead of measured in vitro values of ADME properties as input features. In silico models for nine ADME and physicochemical properties included as features in the rat PK model are frequently updated since these properties are measured for the majority of compounds early in the lead discovery and optimization process. It is likely that the test set compounds for the rat PK model were included in the training sets of in silico ADME models. To ensure that the test set compounds are completely "unseen" by the model, in the second scenario, predictions generated with earlier versions of the in silico ADME models were used instead of in vitro measurements. Earlier ADME models were built before the test set compounds were synthesized and are referred to as "pre-synthesis". The second scenario represents the model predictions for virtual compounds at the point of design.

Figure 4 shows the performance of the model for the default application when in vitro ADME values are used and for the two scenarios. There is a small or no increase in RMSE across all PK endpoints if in silico predictions are used instead of in vitro values as the model input. If pre-synthesis in silico predictions are used, there is an increase in RMSE between 5 and 30% depending on the PK parameter; for example, for the CL endpoint, RMSE = 0.35 for pre-synthesis in silico predictions in comparison with RMSE = 0.27 for in vitro ADME values; for bioavailability F, RMSE = 0.57 and 0.47 for pre-synthesis in silico predictions and in vitro values as inputs, respectively. For V_{ss}, the change in RMSE is very marginal (RMSE = 0.31 and 0.28, respectively, for pre-synthesis in silico predictions and in vitro values as inputs). Thus, the model remains applicable and useful when applied at the point of design, even if predicted compound ADME properties are used as the input. This is of high practical relevance since now some PK parameters in rat models can be predicted with

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

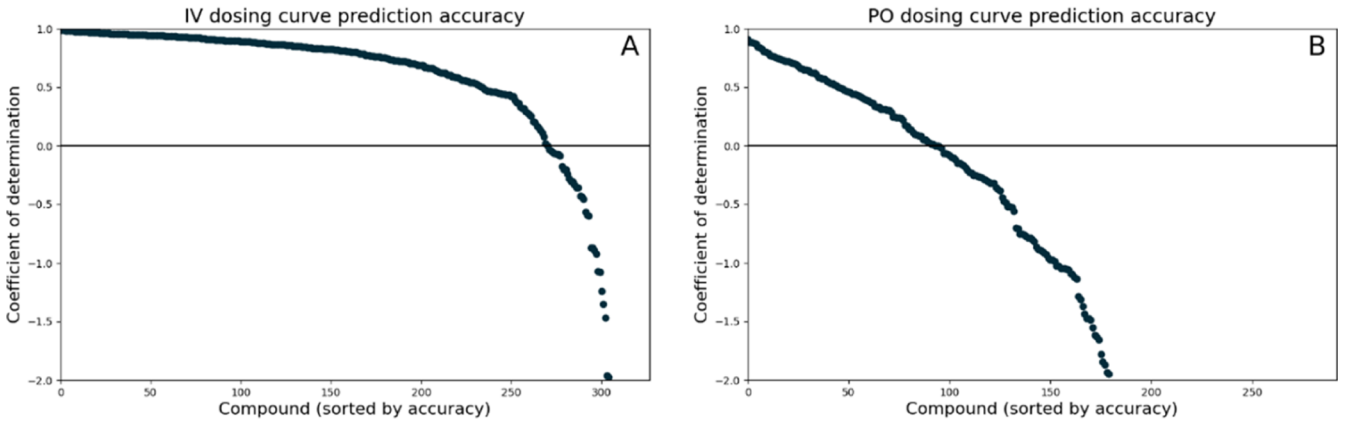


Figure 7. Profiles of accuracy in prediction of PK curves, with R^2 calculated for time curves for which both experimental and modeled values are available, averaged across replicates, for both iv dosing (A) and po dosing (B). Profiles are truncated at $R^2 = -2$.

sufficient accuracy solely based on the chemical structure, without the necessity for experimental measurements.

3.1.6. Confidence in Predictions. A good machine learning model provides an estimation of uncertainty in predictions as well as accurate predictions.[70] The uncertainty quantification can enable detection of out-of-domain examples and identification of less reliable predictions. In this work, the explored algorithms offer three different approaches for the estimation of uncertainty. In the first approach, variability in prediction is captured by generating an ensemble of predictions. This approach is utilized by both deep neural network methods, Alchemite and Chemprop, as well as by GBoost, a decision tree ensemble method. The second approach is inherent in GPR, a Bayesian algorithm that is known to provide a useful quantification of uncertainty.[58,71] The output of GPR is not only a single-point prediction but also a probability distribution where the mean is used as the prediction value and the standard deviation is the estimation of uncertainty. The third approach is the conformal prediction framework,[41,42,72] utilized in the CPSign algorithm based on SVM regression. These three approaches for uncertainty quantification were compared for the example of the CL endpoint. The quality of different uncertainty estimates was evaluated using two metrics: ranking-based confidence curve with the associated quantitative measure AUCO and calibration curve with the associated quantitative measure AUCE. The confidence curves for four different CL models are shown in Figure 5. Clearly, all four confidence curves are far from the perfect "oracle" curve, that is, the ranking order by predicted uncertainty does not correspond to ranking by the real error of prediction. The Chemprop MT method curve, shown in Figure 5B, is closest to the "oracle" curve and provides the best AUCO metric (AUCO = 0.145). Both neural network ensemble methods, Chemprop MT and Alchemite (replace), have better confidence curves than GPR or CPSign methods. The calibration curves for the four CL models are shown in Figure 6. For both neural network methods, the uncertainty in prediction provided by the model significantly underestimated the real uncertainty; corresponding calibration curves are far from perfect calibration. Addition of the aleatoric uncertainty (due to variability in experimental measurements) to the model uncertainty provides a better calibrated model, which is defined by the total uncertainty (see the Methods section and Section 2.4). Both Alchemite (replace) and Chemprop MT benefit from the addition of the experimental uncertainty, as shown in Figure 6A,B, respectively. GPR and CPSign models, on the other hand, produce close to perfect calibration curves, Figure 6C,D. For GPR and CPSign, the addition of the experimental uncertainty was not needed, and the model uncertainty estimation incorporates all sources of uncertainty and represents the total uncertainty. The GPR technique estimates uncertainty using a Bayesian approach and CPSign involves empirical estimation via a conformal prediction framework. The best calibration curve is provided by the GPR model with AUCE = 0.026.

3.2. Models for PK Curve Data. *3.2.1. Accuracy of Curve Prediction.* Profiles of the accuracy in prediction of iv and po concentration−time curves are shown in Figure 7, summarizing the performance of the model on all the test set compounds (312 compounds for iv dosing and 279 for po dosing). Accuracy was evaluated using the coefficient of determination, R^2, between the experimental data and predicted curves across all time points where both the experimental data and predictions were above the limit of detection, averaged over replicates for a given compound. The prediction of iv dosing curves is good, with a median R^2 of 0.82 (median RMSE = 0.41 log units), but the prediction of po dosing curves is poor, with a median R^2 of −0.78 (median RMSE = 0.54 log units). This is likely to be due to po PK being more complex than iv PK because it is strongly influenced by additional mechanisms, such as intestinal absorption and first-pass metabolism. These complex relationships also manifest in higher variability in concentration−time curves and hence a more difficult modeling task. We therefore progress our analysis only of the iv dosing curves.

A set of typical concentration−time curves are shown in Figure 8. Some general trends are noticeable: earlier time points are generally predicted more accurately than later time points, which is likely to be due to more values falling below the measurement tolerance at later times, reducing the amount of precise data for the machine learning model to learn from. The uncertainties on the machine learning predictions are correspondingly greater at later times, providing reassurance that the uncertainty quantification in the model is accurately capturing both this reduction in training data and the increased extrapolation required due to the larger time gaps between measurements at late times.

3.2.2. Calculation of Parameters from Curves. To enable comparison with the results in Section 3.1, PK parameters

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

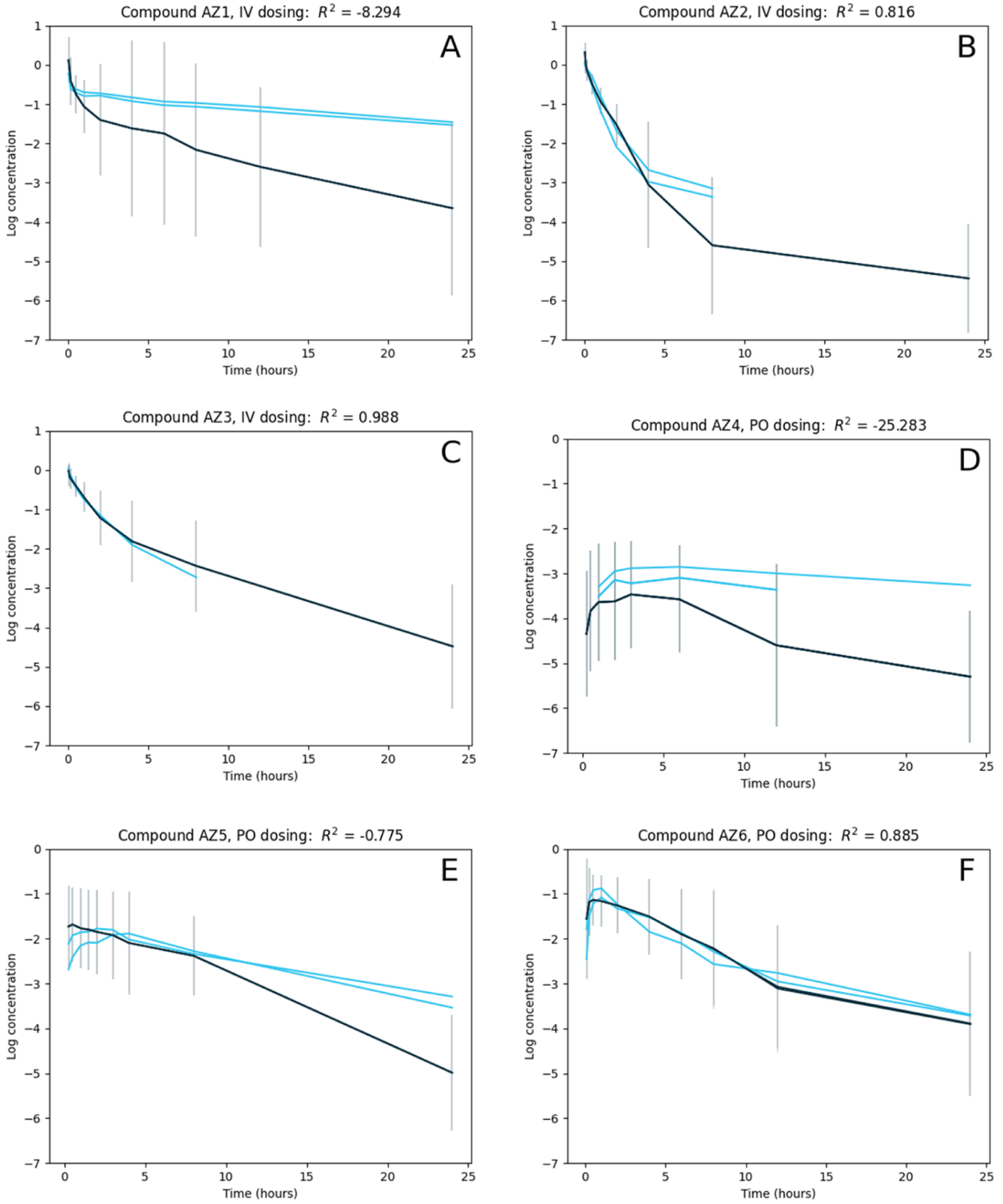


Figure 8. Selection of iv and po dosing curves: experimental data are shown in light blue, including multiple replicates per compound, and the predicted curves are shown in dark blue, with the uncertainty in prediction shown by the vertical gray lines. Coefficient of determination measures for the accuracy of prediction are given in each case. From the top left, these curves show a poorly modeled iv dosing curve (A); an averagely modeled iv dosing curve (B); a well-modeled iv dosing curve (C); a poorly modeled po dosing curve (D); an averagely modeled po dosing curve (E); and a well-modeled po dosing curve (F).

were generated from the predicted curves and compared to the (experimental) PK parameters used for modeling in Section 3.1. These PK parameters had been generated from the true experimental data using a semi-manual process

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

involving cleaning of the underlying data; however, for all PK parameters except V_{ss}, the Pearson correlation between the semi-manual generation and fully automated generation from the raw concentration−time data using MATLAB exceeded 0.97, indicating that the semi-manual process made only small differences to the PK parameter generation. The results for the iv curves are summarized in Table 4 along with the accuracy of the equivalent model predicting the PK parameters directly. AUC and clearance are predicted with equivalent accuracy when generating parameters from the PK curves as when predicting the PK parameters directly, and C_{max} is predicted slightly more accurately when generating parameters from the predicted PK curves, indicating that curve prediction adds value to the analysis of PK. Arbitrary further parameters may also be generated from a predicted curve without requiring training of a new model, in contrast to direct prediction of PK parameters where a new model is required whenever the desired parameters change. Half-life and V_{ss} are predicted less accurately using the curves than when predicted directly: this is likely to be because these parameters are sensitive to the later time behavior of the curve, which, as discussed above, is less accurately captured by the model than the earlier time behavior, and in the case of V_{ss} also due to the difference between the automated and semi-manual methods of generating PK parameters from curve data. These results demonstrate that the machine learning models not only accurately predict the iv curves directly but also the derived PK parameters when a standard PK calculation method is used.

Table 4. Accuracies for PK Parameters Derived From Predictions of iv Curves and Direct Predictions, Using the Alchemite "Replace" Methodology in Both Cases

	generated from predicted curves		directly predicted	
PK parameter	R^2	RMSE	R^2	RMSE
AUC iv	0.54	0.29	0.56	0.29
CL	0.54	0.29	0.57	0.28
C_{max} iv	0.46	0.21	0.42	0.23
$t_{1/2}$ iv	0.30	2.10	0.39	1.93
V_{ss}	0.28	0.34	0.45	0.30

4. CONCLUSIONS

In this work, we built the models for prediction of in vivo rat PK parameters and concentration−time PK profiles from chemical structure representations and experimentally measured ADME properties. We also performed evaluation of multiple machine learning algorithms and approaches to missing data imputation. We observed that models using the "replace" approach generally outperformed models using Alchemite imputation. In silico models trained on a large set of ADME data gave more accurate outcomes than using imputation within a smaller data set. Among the models using the "replace" approach, different machine learning techniques resulted in models of similar accuracy. The neural network algorithms Alchemite and Chemprop yielded the best performing models for the majority of endpoints, with the traditional machine learning algorithms following closely in performance.

Models with acceptable accuracy were achieved for the most important endpoints—clearance (CL), oral bioavailability (F), and volume of distribution (V_{ss}). The model for CL, one of the most important and challenging parameters to optimize in drug discovery, achieved a good performance with R^2 = 0.63 and RMSE = 0.26 (in log units). Furthermore, we benchmarked this model against the WSM, which is routinely applied in decision making for compound prioritization, and showed that the CL model achieved much higher accuracy. Therefore, the CL model provides a useful tool for decision making in early discovery. The model predicts also values higher than the liver blood flow and complements current DMPK tools used for PK prioritization. The model for oral bioavailability achieved R^2 = 0.55 and RMSE = 0.46 (in log units), with RMSE approaching the level of experimental error in the data estimated at 0.43. Overall, sufficiently accurate models were achieved for all the endpoints, except for C_{max} iv, $t_{1/2}$ iv, and $t_{1/2}$ po. We also demonstrated that the models show only a small decrease in accuracy when only in silico data are used; thus, they are useful at the point of design, before compounds are synthesized, and before ADME in vitro properties become available.

In addition to directly predicting in vivo rat PK parameters, we built models of concentration−time profiles, enabling the prediction of concentration scaled by dose at any time point. The accuracy of PK curves prediction for iv dosing is good (the median of individual curve R^2 equals 0.82), but the prediction of curves with po dosing is poor, perhaps due to higher variability in po dosing curve data. PK parameters estimated from predicted iv curves are only slightly less accurate overall than those predicted by the PK models directly, and the curve gives useful additional information.

We have utilized in vivo rat PK parameter predictions as input features to machine learning models for prediction of human PK parameters, AUC po, C_{max} po, and $V_d$ss iv, which we have recently developed.[19] The rat PK parameters are among some of the most important features for the human models.

The models provide a powerful way to guide the design of molecules with optimal rodent PK profiles since they enable the prediction of virtual compounds, and to drive prioritization of compounds for in vivo assays including efficacy experiments. Model usage is expected to reduce the need of animal PK experiments during drug discovery. Furthermore, the developed AI approach is a stepping stone for developing models to predict human PK, ultimately leading to the design of molecules with a desired multiobjective profile early in drug discovery, which will increase efficiency and reduce compound attrition.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.2c00027.

Distributions of the experimental values for PK parameters and of in vitro properties used as input features to the models and detailed performance statistics on developed models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Olga Obrezanova − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca,*

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

Cambridge CB4 0FZ, U.K.; orcid.org/0000-0002-2144-4634; Email: olga.obrezanova@astrazeneca.com

Authors

Anton Martinsson − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg SE-43183, Sweden*

Tom Whitehead − *Intellegens Ltd., Eagle Labs, Cambridge CB4 3AZ, U.K.*

Samar Mahmoud − *Optibrium Ltd., Cambridge CB25 9PB, U.K.*

Andreas Bender − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Cambridge CB4 0FZ, U.K.; Department of Chemistry, Centre for Molecular Informatics, University of Cambridge, Cambridge CB2 1EW, U.K.;* orcid.org/0000-0002-6683-7546

Filip Miljković − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg SE-43183, Sweden;* orcid.org/0000-0001-5365-505X

Piotr Grabowski − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Cambridge CB4 0FZ, U.K.*

Ben Irwin − *Optibrium Ltd., Cambridge CB25 9PB, U.K.*

Ioana Oprisiu − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg SE-43183, Sweden*

Gareth Conduit − *Intellegens Ltd., Eagle Labs, Cambridge CB4 3AZ, U.K.*

Matthew Segall − *Optibrium Ltd., Cambridge CB25 9PB, U.K.;* orcid.org/0000-0002-2105-6535

Graham F. Smith − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Cambridge CB4 0FZ, U.K.;* orcid.org/0000-0003-0393-3650

Beth Williamson − *Drug Metabolism and Pharmacokinetics, Research and Early Development, Oncology R&D, AstraZeneca, Cambridge CB10 1XL, U.K.*

Susanne Winiwarter − *Drug Metabolism and Pharmacokinetics, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), Biopharmaceutical R&D, AstraZeneca, Gothenburg SE-43183, Sweden;* orcid.org/0000-0002-9808-1683

Nigel Greene − *Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, Massachusetts 02451, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.molpharmaceut.2c00027

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

ADME, absorption, distribution, metabolism, and excretion; AI, artificial intelligence; ARD, automatic relevance determination; AUC, area under the concentration−time curve; AUCO, area under the confidence-oracle error; AUCE, area under the calibration error curve; CL, clearance; C_{max}, the maximum plasma concentration; D-MPNN, directed message passage neural network framework; DMSO, dimethyl sulfoxide; EDTA, ethylenediaminetetraacetic acid; *F*, oral bioavailability; GBoost, gradient boosting regression; GPR, Gaussian processes regression; HPMC, hydroxypropyl methylcellulose; iv, intravenous (administration); IVIVE, in vitro to in vivo extrapolation; LC-MS/MS, liquid chromatography−tandem mass spectrometry; LogD, base-10 logarithm of distribution coefficient; LogP, base-10 logarithm of partition coefficient; MT, multitask; NCA, noncompartmental analysis; NN, neural network; PBPK, physiologically based pharmacokinetics; PK, pharmacokinetics; pK_a, negative base-10 logarithm of the acid dissociation constant; po, oral (administration); QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; RBF, radial basis function; ReLU, rectified linear unit; RMSE, root mean squared error; R^2, coefficient of determination; SMILES, simplified molecular input line entry system; ST, single-task; SVM, support vector machine; $t_{1/2}$, half-life; V_d or V_{ss}, volume of distribution; WSM, well-stirred model

■ REFERENCES

(1) Ruiz-Garcia, A.; Bermejo, M.; Moss, A.; Casabo, V. G. Pharmacokinetics in Drug Discovery. *J. Pharm. Sci.* **2008**, *97*, 654−690.

(2) Sturm, N.; Mayr, A.; Le Van, T.; Chupakhin, V.; Ceulemans, H.; Wegner, J.; Golib-Dzib, J.-F.; Jeliazkova, N.; Vandriessche, Y.; Böhm, S.; Cima, V.; Martinovic, J.; Greene, N.; Vander Aa, T.; Ashby, T. J.; Hochreiter, S.; Engkvist, O.; Klambauer, G.; Chen, H. Industry-Scale Application and Evaluation of Deep Learning for Drug Target Prediction. *J. Cheminf.* **2020**, *12*, 26.

(3) Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A. Target Prediction Utilising Negative Bioactivity Data Covering Large Chemical Space. *J. Cheminf.* **2015**, *7*, 51.

(4) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 5918−5922.

(5) Bender, A.; Cortés-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1: Ways to Make an Impact, and Why We Are Not There Yet. *Drug Discovery Today* **2021**, *26*, 511−524.

(6) Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discov. Today* **2021**, *26*, 1040−1052.

(7) Hughes, J.; Rees, S.; Kalindjian, S.; Philpott, K. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239−1249.

(8) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(9) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203−214.

(10) Brian Houston, J. Utility of in Vitro Drug Metabolism Data in Predicting in Vivo Metabolic Clearance. *Biochem. Pharmacol.* **1994**, *47*, 1469−1479.

(11) Sodhi, J. K.; Benet, L. Z. Successful and Unsuccessful Prediction of Human Hepatic Clearance for Lead Optimization. *J. Med. Chem.* **2021**, *64*, 3546−3559.

(12) Sager, J. E.; Yu, J.; Ragueneau-Majlessi, I.; Isoherranen, N. Physiologically Based Pharmacokinetic (PBPK) Modeling and Simulation Approaches: A Systematic Review of Published Models, Applications, and Model Verification. *Drug Metab. Dispos.* **2015**, *43*, 1823−1837.

(13) Rowland, M.; Benet, L. Z.; Graham, G. G. Clearance Concepts in Pharmacokinetics. *J. Pharmacokinet. Biopharm.* **1973**, *1*, 123−136.

(14) Pang, K. S.; Han, Y. R.; Noh, K.; Lee, P. I.; Rowland, M. Hepatic Clearance Concepts and Misconceptions: Why the Well-

Stirred Model Is Still Used Even Though It Is Not Physiologic Reality? *Biochem. Pharmacol.* **2019**, *169*, 113596.
(15) Pang, K. S.; Rowland, M. Hepatic Clearance of Drugs. I. Theoretical Considerations of a "Well-Stirred" Model and a "Parallel Tube" Model. Influence of Hepatic Blood Flow, Plasma and Blood Cell Binding, and the Hepatocellular Enzymatic Activity on Hepatic Drug Clearance. *J. Pharmacokinet. Biopharm.* **1977**, *5*, 625−653.
(16) Davies, M.; Jones, R. D. O.; Grime, K.; Jansson-Löfmark, R.; Fretland, A. J.; Winiwarter, S.; Morgan, P.; McGinnity, D. F. Improving the Accuracy of Predicted Human Pharmacokinetics: Lessons Learned from the AstraZeneca Drug Pipeline Over Two Decades. *Trends Pharmacol. Sci.* **2020**, *41*, 390−408.
(17) Wang, Y.; Liu, H.; Fan, Y.; Chen, X.; Yang, Y.; Zhu, L.; Zhao, J.; Chen, Y.; Zhang, Y. In Silico Prediction of Human Intravenous Pharmacokinetic Parameters with Improved Accuracy. *J. Chem. Inf. Model.* **2019**, *59*, 3968−3980.
(18) Lombardo, F.; Bentzien, J.; Berellini, G.; Muegge, I. In Silico Models of Human PK Parameters. Prediction of Volume of Distribution Using an Extensive Data Set and a Reduced Number of Parameters. *J. Pharm. Sci.* **2021**, *110*, 500−509.
(19) Miljković, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. *Mol. Pharm.* **2021**, *18*, 4520−4530.
(20) Schneckener, S.; Grimbs, S.; Hey, J.; Menz, S.; Osmers, M.; Schaper, S.; Hillisch, A.; Göller, A. H. Prediction of Oral Bioavailability in Rats: Transferring Insights from in Vitro Correlations to (Deep) Machine Learning Models Using in Silico Model Outputs and Chemical Structure Parameters. *J. Chem. Inf. Model.* **2019**, *59*, 4893−4905.
(21) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63*, 8835.
(22) Ye, Z.; Yang, Y.; Li, X.; Cao, D.; Ouyang, D. An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol. Pharm.* **2019**, *16*, 533−541.
(23) Kosugi, Y.; Hosea, N. Direct Comparison of Total Clearance Prediction: Computational Machine Learning Model versus Bottom-Up Approach Using In Vitro Assay. *Mol. Pharm.* **2020**, *17*, 2299−2309.
(24) Kosugi, Y.; Hosea, N. Prediction of Oral Pharmacokinetics Using a Combination of In Silico Descriptors and In Vitro ADME Properties. *Mol. Pharm.* **2021**, *18*, 1071−1079.
(25) Lombardo, F.; Berellini, G.; Obach, R. S. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug Metab. Dispos.* **2018**, *46*, 1466−1477.
(26) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(27) Irwin, B. W. J.; Levell, J. R.; Whitehead, T. M.; Segall, M. D.; Conduit, G. J. Practical Applications of Deep Learning To Impute Heterogeneous Drug Discovery Data. *J. Chem. Inf. Model.* **2020**, *60*, 2848−2857.
(28) Winiwarter, S.; Middleton, B.; Jones, B.; Courtney, P.; Lindmark, B.; Page, K. M.; Clark, A.; Landqvist, C. Time Dependent Analysis of Assay Comparability: A Novel Approach to Understand Intra- and Inter-Site Variability over Time. *J. Comput. Aided Mol. Des.* **2015**, *29*, 795−807.
(29) Wenlock, M. C.; Potter, T.; Barton, P.; Austin, R. P. A Method for Measuring the Lipophilicity of Compounds in Mixtures of 10. *J. Biomol. Screen* **2011**, *16*, 348−355.
(30) Wan, H.; Holmen, A. High Throughput Screening of Physicochemical Properties and in Vitro ADME Profiling in Drug Discovery. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 315−329.
(31) Fredlund, L.; Winiwarter, S.; Hilgendorf, C. In Vitro Intrinsic Permeability: A Transporter-Independent Measure of Caco-2 Cell Permeability in Drug Design and Development. *Mol. Pharm.* **2017**, *14*, 1601−1609.
(32) Sohlenius-Sternbeck, A.-K.; Afzelius, L.; Prusis, P.; Neelissen, J.; Hoogstraate, J.; Johansson, J.; Floby, E.; Bengtsson, A.; Gissberg, O.; Sternbeck, J.; Petersson, C. Evaluation of the Human Prediction of Clearance from Hepatocyte and Microsome Intrinsic Clearance for 52 Drug Compounds. *Xenobiotica* **2010**, *40*, 637−649.
(33) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 125−134.
(34) Temesi, D. G.; Martin, S.; Smith, R.; Jones, C.; Middleton, B. High-Throughput Metabolic Stability Studies in Drug Discovery by Orthogonal Acceleration Time-of-Flight (OATOF) with Analogue-to-Digital Signal Capture (ADC). *Rapid Commun. Mass Spectrom.* **2010**, *24*, 1730−1736.
(35) Wan, H.; Bergström, F. High Throughput Screening of Drug-Protein Binding in Drug Discovery. *J. Liq. Chromatogr. Relat. Technol.* **2007**, *30*, 681−700.
(36) Waters, N. J.; Jones, R.; Williams, G.; Sohal, B. Validation of a Rapid Equilibrium Dialysis Approach for the Measurement of Plasma Protein Binding. *J. Pharm. Sci.* **2008**, *97*, 4586−4595.
(37) Austin, R. P.; Barton, P.; Mohmed, S.; Riley, R. J. The Binding of Drugs to Hepatocytes and Its Relationship to Physicochemical Properties. *Drug Metab. Dispos.* **2005**, *33*, 419−425.
(38) Olsson, T.; Sherbukhin, V. *SELMA, Synthesis and Structure Administration (SaSA); Documentation*; AstraZeneca R&D: Mölndal, Sweden, 2002.
(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(40) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.
(41) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54*, 1596−1603.
(42) CPSign; Documentation; https://arosbio.com/cpsign/docs/latest/ (accessed 12 Aug, 2020).
(43) Alvarsson, J.; Eklund, M.; Andersson, C.; Carlsson, L.; Spjuth, O.; Wikberg, J. E. S. Benchmarking Study of Parameter Variation When Using Signature Fingerprints Together with Support Vector Machines. *J. Chem. Inf. Model.* **2014**, *54*, 3211−3217.
(44) Oprisiu, I.; Winiwarter, S. In silico ADME modelling. In *Systems Medicine: Integrative, Qualitative and Computational Approaches*; Wolkenhauer, O., Ed.; Elsevier Inc., 2020; Vol. *2*, pp 208−222.
(45) Whitehead, T. M.; Irwin, B. W. J.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1197−1204.
(46) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.
(47) *ACD/Labs Software*; Advanced Chemistry Development, Inc.: Toronto, Ontario, Canada, 2015.
(48) *ClogP*, version 4.3; Pomona College and BioByte, Inc.: Claremont, CA, US.
(49) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.
(50) Wood, D. J.; Buttar, D.; Cumming, J. G.; Davis, A. M.; Norinder, U.; Rodgers, S. L. Automated QSAR with a Hierarchy of Global and Local Models. *Mol. Inf.* **2011**, *30*, 960−972.
(51) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*; Cortes, C.,

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX

Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; Vol. *28*.
(52) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595−608.
(53) StarDrop, version 6.5; Optibrium Ltd.: Cambridge, U.K.; https://www.optibrium.com/stardrop (accessed 27 Aug, 2021).
(54) SMARTS-A Language for Describing Molecular Patterns. Documentation; Daylight Chemical Information Systems Inc.: Laguna Niguel, CA, USA; https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 27 Aug, 2021).
(55) Landrum, G. RDKit: Open-Source Cheminformatics; Documentation; https://www.rdkit.org/docs/ (accessed 25 Nov, 2021).
(56) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, 2010; pp 807−814.
(57) MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, United Kingdom, 2003.
(58) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, 2006.
(59) Obrezanova, O.; Csányi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.
(60) Obrezanova, O.; Gola, J. M. R.; Champness, E. J.; Segall, M. D. Automatic QSAR Modeling of ADME Properties: Blood-Brain Barrier Penetration and Aqueous Solubility. *J. Comput. Aided Mol. Des.* **2008**, *22*, 431−440.
(61) Burden, F. R. Quantitative Structure−Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830−835.
(62) *Matlab*, version 9.6.0.1072779 (R2019a); The Mathworks, Inc.: Natick, Massachusetts, 2019; https://www.mathworks.com/.
(63) Conduit, B. D.; Jones, N. G.; Stone, H. J.; Conduit, G. J. Design of a Nickel-Base Superalloy Using a Neural Network. *Mater. Des.* **2017**, *131*, 358−365.
(64) Bergstra, J.; Yamins, D.; Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *Proceedings of the 30th International Conference on Machine Learning, June 2013*; Dasgupta, S.; McAllester, D., Eds.; PMLR, 2013; Vol. *28*, pp 115−123.
(65) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697−2717.
(66) Cerny, M. A. Prevalence of Non-Cytochrome P450-Mediated Metabolism in Food and Drug Administration-Approved Oral and Intravenous Drugs: 2006-2015. *Drug Metab. Dispos.* **2016**, *44*, 1246−1252.
(67) Williamson, B.; Colclough, N.; Fretland, A. J.; Jones, B. C.; Jones, R. D. O.; McGinnity, D. F. Further Considerations Towards an Effective and Efficient Oncology Drug Discovery DMPK Strategy. *Curr. Drug Metab.* **2020**, *21*, 145−162.
(68) Riley, R. J.; McGinnity, D. F.; Austin, R. P. A Unified Model for Predicting Human Hepatic, Metabolic Clearance from in Vitro Intrinsic Clearance Data in Hepatocytes and Microsomes. *Drug Metab. Dispos.* **2005**, *33*, 1304−1311.
(69) Noncompartmental analysis; Documentation; The Mathworks, Inc.: Natick, Massachusetts; https://uk.mathworks.com/help/simbio/ug/non-compartmental-analysis.html (accessed 25 Nov, 2021).
(70) Mervin, L. H.; Johansson, S.; Semenova, E.; Giblin, K. A.; Engkvist, O. Uncertainty Quantification in Drug Design. *Drug Discov. Today* **2021**, *26*, 474−489.
(71) Hie, B.; Bryson, B. D.; Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Syst.* **2020**, *11*, 461−477.
(72) Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1269−1281.

https://doi.org/10.1021/acs.molpharmaceut.2c00027
Mol. Pharmaceutics XXXX, XXX, XXX−XXX