Prediction of compound *in vivo* pharmacokinetics in rats using machine and deep learning

Olga Obrezanova[1], Anton Martinsson[2], Tom Whitehead[3], Samar Mahmood[4], Andreas Bender[1], Filip Miljković[2], Piotr Grabowski[1], Ben Irwin[4], Ioana Oprisiu[2], Gareth Conduit[3], Matthew Segall[4], Graham Smith[1], Beth Williamson[5], Susanne Winiwarter[6], and Nigel Greene[7]

[1]*Data Science and AI, Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca. Cambridge, UK*
[2]*Data Science and AI, Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden*
[3]*Intellengens, Eagle Labs, Chesterton Road, Cambridge, UK*
[4]*Optibrium,*
[5]*Drug Metabolism and Pharmacokinetics, Research and Early Development, Oncology R&D, AstraZeneca, Cambridge, UK*
[6]*Drug Metabolism and Pharmacokinetics, Early CVRM, Biopharmaceutical R&D, AstraZeneca, Gothenburg, Sweden*
[7]*Data Science and AI, Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, US*

ABSTRACT

Rodent pharmacokinetic (PK) data and human and animal *in vitro* systems are utilised 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 human. In addition, prediction of *in vivo* properties can be used to improve design during drug discovery, help to select compounds with better properties, and reduce the number of *in vivo* experiments. In this study, we build machine learning models able to predict rat *in vivo* pharmacokinetic parameters, including rat oral bioavailability and clearance, which utilise molecular chemical structure and, either measured or predicted, *in vitro* ADME parameters. The models were trained on internal *in vivo* rat PK data for over 3,000 diverse compounds from multiple projects and therapeutic areas. We compare performance of various traditional machine learning algorithms and deep learning approaches, including graph convolutional neural networks that encode molecule graph structure. The best models achieved R^2=0.63 for clearance and R^2=0.55 for bioavailability. The models provide a powerful way to guide the design of molecules with optimal PK profiles, to enable the prediction of virtual compounds e.g. during DMTA cycles, and to drive prioritisation of compounds for *in vivo* assays.

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 – *i.e.*, 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 modelling *in vivo* relevant properties, such as PK, is much less pronounced at this stage. One reason is that domains differ significantly with respect to the data

available:[5,6] In some areas *in vitro* proxy assays can be run to characterize compounds,[7] such as biochemical assays, or also assays for PK-*related* properties such as logD or solubility, which give rise to large numbers of available data points, generated in a relatively consistent manner. This renders this – proxy – space for drug discovery relatively amenable to current developments in the machine learning domain, such as deep learning.[8] Regarding *in* vivo pharmacokinetics data (as well as *in vivo* data more generally), however, data generation is more costly and complex, leading generally to a lack of data in this domain, which does not render the application of some algorithms as straightforward.[6] On the other hand, due to the direct therapeutic relevance of *in vivo* assays, as well as their high cost, modelling this type of endpoints equally provides a *stronger incentive* to provide *in silico* models in this area, provided the hurdle of access to suitable data and its normalization for model generation can be overcome. Case in point, it has been shown that failure rates *in the clinical phases* are what makes drug discovery (and its failures) so costly[5,9] – and hence the more we are able to anticipate compounds behaviour *in man* (as opposed to *in vitro* assays) early on, the more impact any assays will have on overall project success[10] when deciding which compounds to take forward in a given project. Provided *in silico* models are able to model endpoints which are relevant for compound behaviour in man, they are able to hence support such decisions.

The purpose of this study is to develop machine learning models for prediction of *in vivo* rat PK parameters utilising molecular chemical structure and *in vitro* measured (or predicted) ADME and physicochemical properties of compounds. The model is based on a dataset of more than 3,000 *in vivo* studies with intravenous (iv) and oral (po) administration with a range of PK endpoints, including the area under the concentration-time curve (AUC), the maximum plasma concentration (C_{max}), half-life ($t_{1/2}$), clearance (CL), volume of distribution (V_{ss}), and oral bioavailability (F) as well as concentration-time profiles.

In a drug discovery project these parameters, depending on the particular indication and compound, need to be within certain ranges for the drug to achieve 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 to achieve a therapeutic efficacy at the site of action, while clearance needs to be sufficiently low due to the same reason as well as to achieve practically feasible dosing regimes. At the same time, a balance between efficacy and safety needs to be found. Here, the key parameter related to safety is C_{max} which needs to be generally below the Maximum Tolerated Dose (MTC), while at the same time the concentration needs to be higher than the Minimum Effective Concentration (MEC). While details differ significantly from case to case, the PK profile of a compound is as much a requirement to achieve efficacy and safety of a compound *in vivo* as its intrinsic properties, and hence the computational prediction of compound concentration over time is of crucial importance for compound selection. This is the case for both individual projects, as well as for computational approaches such as those involving the Design-Make-Test-Analyse (DMTA) cycle,[11] and in particular with the view of eventual efficacy and safety of the compound in the clinic in mind.

Current approaches to *in* vivo PK prediction are (among others) the well-stirred model,[12,13] and physiologically-based pharmacokinetics (PBPK) modelling.[14] The well-stirred models assumes the drug concentration in liver to be equal to that of incoming blood ('well stirred'), and it firstly comprises the generation of *in* vitro data, in particular those on either liver microsomes and/or hepatocytes, and a subsequent extrapolation step to humans.[13] Significant advances have been made recently in anticipating human *in* vivo PK from *in* vitro data, and recent work from AstraZeneca[15] 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 two-fold". This also underpins in particular the 'right safety' aspect of drug discovery at AstraZeneca, as described in the '5R' strategy[16] (with the other 'R's being right target, right tissue, right patient, and right commercial potential) which has increased success rates from candidate drug nomination to phrase III competition from 4% between 2005 and 2010 to 19% between 2012 and 2016.

Differences between compound behaviour in different types of cells to determine their *in vitro* behaviour exist, and they have recently been better understood.[17] Hence, overall, *in vitro* anticipation of compound *in vivo* PK provides a cornerstone of compound evaluation at early preclinical phases currently. PBPK models on the other hand are usually applied later in drug discovery projects, and they comprise an approximation of the (physiological) human body and its major organs, and modelling compound concentration in different organs as a series of coupled differential equations which need to be parameterized in the first place.[14] On the one hand, this approach – where successful – is able to provide insights into compound exposure in different (major) organs as a function of time, which is physiologically of tremendous value (given that accumulation may occur, leading to local concentrations which differ from those observed in plasma). Also, the influence of intrinsic factors, such as sex and age, as well as extrinsic factors (such as drug-drug interactions) on exposure can be modelled. On the other hand, the parametrization of PBPK models requires compound-dependent parameters and manual input into model development, which is hence generally not possible for compounds in a high-throughput manner.

We can conclude from the above discussion that both the well-stirred model and PBPK modelling approaches require the experimental determination of *in vitro* parameters and subsequent extrapolation/modelling, which is hence not an approach which is feasible purely based on chemical structure. This, however, is what would be required to also assess *virtual* compounds, be it during the design process in a drug discovery project, or *e.g.* in the context of generative models for prioritizing large numbers of structures *in silico*.

For practical purposes, hence the prediction of PK parameters *directly based on chemical structure* would be desirable, both for individual project use and in the context of current computational approaches, such as DMTA cycles, in order to move compound prioritization from proxy properties, such as bioactivity on target and a series of *in vitro* properties, to more relevant *in vivo* space.[5,6]

Returning to the *in silico* modelling of compound *in vivo* PK directly based on chemical structure recent approaches shall now be briefly summarized here. In one of the first studies of its type Lowe *et al.*[18] modelled rat and human microsomal intrinsic clearance, as well as plasma protein binding represented as the fraction of compound unbound, using Artificial Neural Networks, Supports Vector Machines and other approaches in combinations with 2D and 3D descriptors for 400-600 compounds per endpoint which was compiled from literature. Both human and rat clearance models were able to capture trends in the data rather well, while models for fraction unbound were based on unbalanced datasets (for compounds with generally low fraction unbound) and its practical utility is less easy to assess. A subsequent study[19] established QSPR models for four human pharmacokinetic parameters, including volume of distribution at steady state, clearance half-life, and fraction unbound in plasma, using a data set consisting of 1,352 drugs (which is currently also the largest publicly available dataset of its type[20]). For clearance endpoint 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 was modelled in this work, as well as due to a bias towards compounds with low clearance, due to the way the dataset was derived. Also more specific models for volume of distribution have been described recently,[21] based on Random Forest methods, and evaluated using an independent test

set of 213 compounds, which was found to compare favourably to methods based on *in vitro* properties.

Other studies compared *in silico* predictive models for PK with PBPK, here considering also different tissues, on 159 structurally varied types of drugs, food components, and industrial chemicals.[22] In this comparison, an *in silico* one-compartment model and a PBPK model comprising the gut, liver, main, and kidney compartments were developed in parallel. Compounds were 'virtually dosed' orally in rats, and the relationship between the simulated internal concentrations in tissue/plasma and their lowest-observed-effect levels was determined. It was found that the Cmax and AUC obtained by one-compartment models and modified simple PBPK models were closely correlated. While this work is conceptually different from modelling PK properties of compounds solely based on chemical structure it should still be mentioned here, since, in combination with bioactivity/assay endpoints of relevance for toxicity in a particular organ, this direction of work may well be suitable to move the field towards organ-based risk assessment.[23,24]

Recently also deep learning and graph convolutional algorithms have been applied to *in vivo* PK modelling. In a study on a large dataset of about 1,900 *in vivo* datapoints[25] researchers at Bayer modelled intravenous (*i.v.*) and oral drug exposure and oral bioavailability in rats using a variety of hybrid modelling approaches, such as using different transformations (such as deep neural networks) and different types of modelling methods. Compounds were described as either (a) six *experimentally* determined in vitro and physicochemical properties, namely, membrane permeation, free fraction, metabolic stability, solubility, pKa value, and lipophilicity; or (b) *the outputs of six in silico* absorption, distribution, metabolism, and excretion models trained on the same properties; or (c) the chemical structure encoded as fingerprints or SMILES strings. The authors found that exposure after iv administration can be predicted similarly well using experimental and predicted properties as input. The model errors for exposure after oral 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 datasets from Merck the authors of another study[26] 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. Imputation has also recently shown to improve performance on a wide variety of ADME endpoints.[27] Using transfer learning and multitask learning[28] one recent model was pre-trained on over 30 million bioactivity data points, and then four human pharmacokinetic parameters for 1104 FDA approved small molecule drugs were modelled, namely oral bioavailability, plasma protein binding, apparent volume of distribution at steady-state and elimination half-life. The multitask learning model generally has shown best performance for the endpoints modelled, although not with a very large margin in some cases.

While studies using machine learning for PK prediction exist, one key question is whether they perform better with respect to predictive power for the *in vivo* situation than extrapolating from *in vitro* data. In this regard, a recent study compared the *in vitro* to *in vivo* extrapolation (IVIVE) approach and machine learning approaches for *in vivo* clearance prediction in rat[29] on a structurally diverse set of 1,114 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 well-stirred model was found to perform substantially worse

compared to machine learning approaches. Similar conclusions were made in a study by the same authors which compared machine learning models for *in vivo* AUC after oral administration to IVIVE approach using a dataset of 595 compounds.[30] Both of these studies, in agreement with our findings and the current work, suggest that *in silico* machine learning models for compound *in vivo* PK properties are of practical value for compound prioritization.

From the above survey we can conclude that there exists prior art in the area of modelling *in vivo* compound PK based on chemical structure. Some endpoints, such as volume of distribution have been shown before to be modellable 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 datasets, and/or the number of PK endpoints modelled was limited to a small number of them.

To address this point, in this work we will describe a machine learning model that predicts a wide range of rat *in vivo* PK parameters, bioavailability (F), clearance (CL), volume of distribution (V_{ss}), AUC, C_{max} and $t_{½}$. The model is trained and validated on a large dataset of more than 3,000 compounds. The combination of endpoints modelled and the number of *in vivo* data used for training, to the best of the knowledge of the authors, makes it the most comprehensive model of its type, both in output property space, and with respect to chemical space coverage. Furthermore, given its non-clinical nature, the datasets used span property ranges of beyond those of just successful drugs, and hence better model predictivity across the value range can be expected from this data set. We explored state-of the-art AI approaches, such as graph convolutional neural networks that encode molecule chemical graph structure,[31] as well as traditional machine learning algorithms utilising 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 utilising corresponding *in silico* predictions or using deep learning technology[27] able to handle sparse and noisy experimental data, were explored. The model is based on input properties which can be predicted *in silico*, and hence it can be applied to any compound structure, including virtual compounds, guiding design of compounds with optimal PK profiles. The model can be used to drive prioritisation of compounds for *in vivo* assays and to inform compound selection in DMTA cycles, which will increase efficiency of the drug discovery and reduce compound attrition.

2. METHODS

2.1. Data set. *In vivo* rat PK data (intravenous (iv) and oral (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 dataset focused on low dose PK studies e.g. 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 modelling; five parameters corresponded to the iv route: the area under the concentration curve (AUC iv), the maximum plasma concentration (C_{max} iv), half-life ($t_{1/2}$ iv), clearance (CL) and volume of distribution at steady state (V_{ss}), and four parameters corresponded to the po route: AUC po, C_{max} po, $t_{1/2}$ po and oral bioavailability (F), defined as the percentage of a po dose that reaches the systemic circulation given by the following equation:

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

Where, D_{iv} and D_{po} are iv and po administration doses, respectively.

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 5 compounds at low doses (see above). Standard formulations for iv were solutions containing cyclodextrin or other solubilizing agents in acceptable quantities, whereas for po, suspensions using hydroxypropyl methylcellulose (HPMC) were usually preferred. Blood samples were taken at pre-defined timepoints post dosing, usually 10 occasions up to 24 h, collected in EDTA-containing tubes and centrifuged at 4000 g for 5 min at 4 °C to obtain plasma. Plasma samples were stored at -75 °C until they were analysed using a liquid chromatography-tandem mass spectrometry (LC-MS/MS). The resulting time-concentration profiles were evaluated using non-compartmental analysis (NCA).

2.1.2. Data curation. The AUC (uM*h), C_{max} (uM) 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 molecular weight higher than 750 Da were excluded from the dataset. The final dataset consisted of 3070 compounds.

2.1.3. Data transformations. AUC (iv and po), C_{max} (iv and po), 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 cut-off value a_{min} was defined in 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, a_{min}=-2 for V_{ss}. No transformation was applied to half-life (h) iv and po. F was first normalised by the maximum value in the dataset (F = 160 %), normalised values below 0.01 were set to 0.01, then the logit transformation was used, where logit y = log_{10} (y/(1-y)). Concentration values in time-concentration profiles were log_{10}-transformed, no a_{min} cut-off was applied.

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.

2.1.5. In vitro ADME properties. Nine experimentally obtained ADME and physicochemical properties were added to the dataset 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 performed early in lead optimisation. The properties describe compound lipophilicity, solubility, permeability, intrinsic metabolic clearance and plasma protein and hepatocyte binding:

- LogD
- Solubility (Dried 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 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 dataset. The assay-dependent percentage of missing values ranged from 6% (LogD) to 55% (fraction unbound in rat hepatocytes).

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.[32,33] Solubility was measured as thermodynamic solubility from DMSO stock solution, where DMSO was evaporated before analysis again using a shake-flask method.[32,34] Caco-2 intrinsic permeability was measured in the presence of a transporter inhibitor cocktail considering a pH gradient using pH 6.5 at the apical side and pH 7.4 at the basolateral side, whereas pH was 7.4 on both sides when measuring the Caco-2 efflux ratio.[35] 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.[36-38] Plasma protein binding data was generated using equilibrium dialysis.[37,39,40] Fraction unbound in rat hepatocytes was also determined using equilibrium dialysis.[41]

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 dataset. The models for these properties were developed using large internal datasets (≥4000 compounds in smaller datasets and up to 160,000 compounds in the larger datasets). Models for the Caco-2 intrinsic permeability and Caco-2 efflux ratio were developed using the random forest algorithm with OESelma molecular property descriptors[42] (see section 2.2.1). Scikit-learn implementation was used for the random forest.**Error! Reference source not found.** The rest of the properties were modelled using a support vector machine with signature descriptors[44] and the conformal prediction framework[45] implemented in the CPSign software.[46,47] A temporal test set (10% of the data) was used for validation, where a dataset was split chronologically into the training and test sets and 10% of latest data are reserved for the test set. The approach represents real life scenario of model usage. The models are regularly updated, with frequency of update varying between 1 to 6 months, depending on the amount data being generated for each property. Model performance is monitored continuously by predicting the new data before each model update. Details of models performance and methods were described recently by Oprisiu and Winiwarter.[48]

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 'replace' approach, was to replace missing *in vitro* values with corresponding *in silico* predictions. High correlation was observed between experimentally measured values for properties and corresponding predictions (correlation coefficient in the range 0.80-0.95) which is not surprising because the experimental data is likely to be contained within training sets of the *in silico* models. The second approach was an imputation approach built-in within Alchemite method,[27,49] referred to as 'impute' approach, it is described below in section 2.3.5.

2.1.9. Training/test data set split. Temporal set split was used to divide the data into the training and test sets, that is around 10% of compounds (312 compounds) with latest synthesis date were separated into the test set. The test set was not used during training and hyperparameter optimisation. Table 1 describes number of compounds in the training and test set for all endpoints.

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 (multi-task format)	2758	312
	Non-aggregated format	
Overall (multi-task format)	9923	1183
Concentration of dose-time profile iv	5895	632
Concentration of dose-time profile po	4266	578

2.2. Chemical descriptors.

2.2.1. OESelma molecular properties. The OESelma descriptors were generated by AstraZeneca's in-house program OESelma.[42] They comprise around 100 common 1D and 2D molecular descriptors related to physico-chemical properties, such as size, ring structure, flexibility, atom types, hydrogen bonds, polarity, electronic environment, partial atom charge, and lipophilicity, including connectivity indices.[50] Additionally, logD and logP from ACDLabs[51] and logP from Biobyte[52] were included in the descriptor set. These descriptors have been shown useful in QSAR modelling, see e.g. works by Bruneau,[53] Wood *et al.*[54] and Fredlund *et al*.[35]

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.[55,56] In this study, the directed message passage neural network framework (D-MPNN) Chemprop[31] 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, atomic mass) as nodes and bond features (bond type, conjugation, ring membership, 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 neighbouring atoms 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.[31]

2.2.3. Signature descriptors. Signatures are 2D descriptors[44] which represent 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[46] with default settings. Signature heights were ranging from 0 to 3.

2.2.4. StarDrop descriptors. The descriptors were calculated with the Auto-Modeller™ module of the StarDrop™ software[57] 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 defined by SMARTS strings.[58]

2.3. Description of modelling techniques

2.3.1. Chemprop. The readout phase of Chemprop is a feed-forward neural network.[31] Five-fold cross-validation based on scaffold splits was performed 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.[59] As a result, the cross-validation performance is based on 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.[60] 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 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. Gaussian Processes is a kernel-based Bayesian probabilistic method[61,62] which was previously successfully utilised for ADME and PK modelling.[63-65,29,30] Matlab 2019a implementation was used in this work.[Error! Reference source not found.] Five kernel functions were explored: exponential, squared exponential, rational quadratic, ARD (Automatic Relevance Determination) squared exponential and ARD exponential. For the rest of the hyperparameters, the defaults were accepted. Ten-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 which 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 gradient descent like procedure. Gradient boosting regression as implemented within Scikit-learn was used.[Error! Reference source not found.] 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. SVM – CPSign. CPSign algorithm[46] is a support vector machine with signature descriptors[44] and a conformal prediction framework.[45] 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 has been used in a variety of chemistry and materials science domains.[27,49,67] In this work it was used to predict either pharmacokinetic parameters, in common with the methods described above, or pharmacokinetic 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 using three different classes of input data: the 'iv' approach used only *in vitro* data as input, which was 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 input, relying on Alchemite to identify the correlations between the datasets 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.[68]

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 modelled simultaneously, using the 'replace' approach to deal with missing *in vitro* data. Alchemite uses the measurement time as an additional input when modelling 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. At 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 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 modelling purposes these points were replaced by the minimum measured concentration in the dataset (4.3×10^{-6} µM/µmol/kg) to ensure the model was aware of the tendency to low concentrations at late times. The log-concentration was modelled to provide accuracy over multiple order of magnitude of concentration.

2.3.6. Combinations of algorithms and descriptors. Not all combinations of descriptors and modelling techniques were investigated. Table 2 describes approaches and algorithms which were explored and specifies abbreviations used for various techniques.

Table 2: Combinations of features, algorithms and approaches explored together with respective abbreviations.

Algorithm		ChemProp Multi-Task	ChemProp Single-Task	Gradient Boosting Regression	Gaussian Processes	Support Vector Machine	Alchemite Multi-Task
Input features		Graph con-volutions ADME properties	Graph con-volutions ADME properties	 OESelma descriptors ADME properties	 OESelma descriptors ADME properties	Signature descriptors OESelma descriptors ADME properties	 StarDrop descriptors ADME properties
Use of ADME features and missing values approaches		*Aggregated format*					
	ADME *in vitro* ('replace' for missing values)	ChemProp MT	ChemProp ST	GBoost	GPR	CPSign	Alchemite (replace)
	ADME *in vitro* ('impute' for missing values)						Alchemite (iv)
	ADME *in vitro* ('impute' for missing values) + ADME *in silico*						Alchemite (ivis)
	ADME *in silico*			GBoost (is)			
		Non-aggregated format					
	ADME *in vitro* ('replace' for missing values)						Alchemite (replace) nAgg
	ADME *in vitro* ('impute' for missing values)						Alchemite (iv) nAgg
	ADME *in vitro* ('impute' for missing values) + ADME *in silico*						Alchemite (ivis) nAgg

2.4. Evaluation of uncertainty estimates. Two metrics were considered to evaluate the quality of different uncertainty estimates – ranking-based and calibration-based.[69]

2.4.1. Ranking-based confidence curve. To construct the confidence curve, the compounds are ordered by the predicted uncertainty in a decreasing order. The compounds with highest uncertainty are gradually removed and RMSE is measured for a remaining subset. RMSE of the subset (100-n % of compounds with the lowest uncertainty) is plotted as a function of confidence percentile n.[69] 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 (AUCO) error which is defined as the difference between the areas under the both curves, can be used as a quality metric.

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, 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, i.e. 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 (AUCE) curve which is defined as the absolute difference between the areas under the calibration and perfect curves, can be used as a quality metric.[69]

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 an experimental error.

2.5. Description of well-stirred model. Hepatic elimination remains the primary route of elimination for drugs[70] hence to understand if hepatic metabolic enzymes present in hepatocytes or liver microsomes can account for the CL of a compound in animals, *in vitro* and *in vivo* extrapolation (IVIVE) using the well stirred model (WSM) is routinely applied.[71-73] 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 human.

2.6. **Calculation of PK parameters from predicted concentration-time profiles.** PK parameters were calculated from predicted concentration-time curves by the Noncompartmental analysis (NCA) using SimBiology App of Matlab R2019a.[74,75] Predicted values that fell below half the minimum of experimentally observed value (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 DISCUSSIONS

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 modelling techniques were investigated (described in section 2.3.6, Table 2).

The results of modelling efforts for the aggregated data format are summarised in Figure 1 showing coefficient of determination (R^2) evaluated on the test set. The detailed results including RMSE on the test set are shown in Figure S1 and Table S1. Models with good accuracy were achieved for the majority of the endpoints, except for C_{max} iv, $t_{1/2}$ iv and $t_{1/2}$ po. Figure 1 also shows that there is no single technique which exceeds other methods across all endpoints. Alchemite (iv) and Alchemite (ivis) use Alchemite method of imputation, based respectively on *in vitro* data only or *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 show that the models using the 'replace' approach generally outperform models using imputation. Performance of the Alchemite (ivis) models, which uses 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 coefficients of determination 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 to 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 is not shown here). Alchemite (iv) represents 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 dataset, which aligns with our expectations.

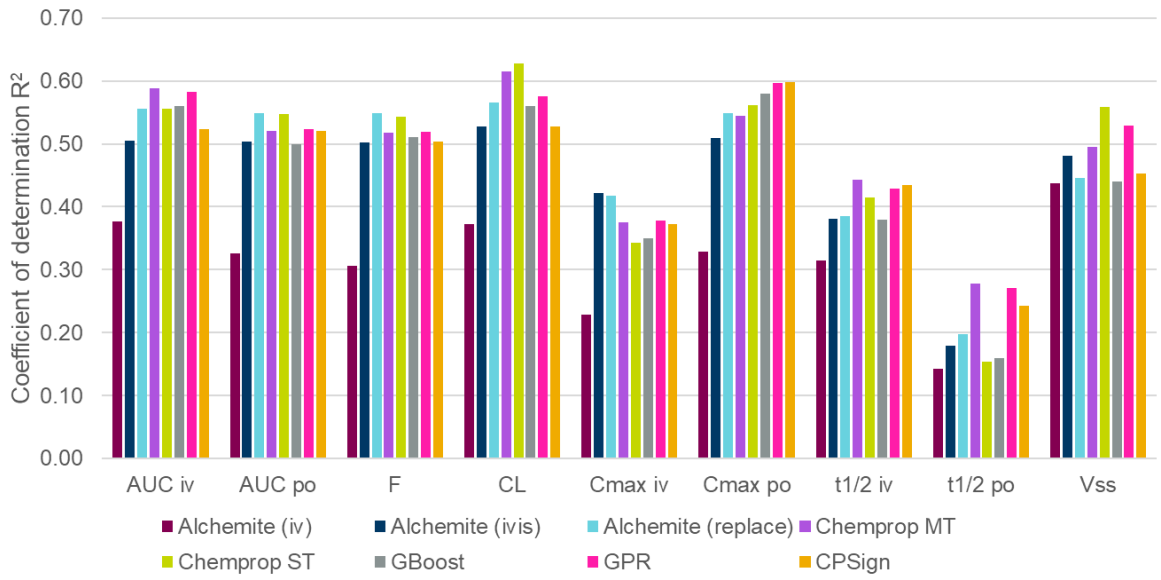

Figure 1: Coefficient of determination (R^2) on the test set for the nine PK parameters using different models built using aggregated data format. Alchemite (iv) is Alchemite multi-task DNN algorithm with in vitro features and imputation (plum bars), Alchemite (ivis) is Alchemite algorithm with in silico 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 Alchemite algorithm (light blue bars), Chemprop MT and Chemprop ST

are Chemprop DNN in multi-task and single-task modes (purple and green bars, respectively), GBoost is a Gradient Boosting Regression (grey bars), GPR is a Gaussian Processes regression (pink bars) and CPSign is a Support Vector Machine Conformal Regression (orange bars).

Focussing on the models using the 'replace' approach, for the majority of endpoints neural network algorithms, Alchemite (replace), Chemprop MT and ST, yield the best performing model with the exception of C_{max} po where Gaussian Processes model (GPR) provides the best performance. The single task neural network models provide broadly equivalent performance to multi-task models on most of the endpoints, for V_{ss} the Chemprop single task model performed better than others. The traditional machine learning algorithms, Gaussian Processes and Gradient Boosting closely follow neural network models in performance for most of endpoints. SVM with 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 – 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 Gaussian Processes due to a different initialisation of weights in NN methods and different (random) cross-validation splits which would in turn affect hyperparameter optimisation. Due to this variability, which was not fully captured, the performance of all 'replace' algorithms apart from 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 are shown in Table 3. The models where difference between RMSE and the lowest RMSE did not exceed 0.005, were considered of similar performance.

Table 3: The best model for each PK parameter together with coefficient of determination (R^2) and RMSE.

PK parameter	Best model(s)	R^2	RMSE
AUC iv	Chemprop MT = GPR	0.59	0.28
AUC po	Alchemite (replace) = Chemprop ST	0.55	0.61
F	Alchemite (replace) = Chemprop ST	0.55	0.46
CL	Chemprop ST = Chemprop MT	0.63	0.26
C_{max} iv	Alchemite (ivis) = Alchemite (replace)	0.42	0.22
C_{max} po	GPR = CPSign	0.60	0.56
$t_{1/2}$ iv	Chemprop MT	0.44	1.84
$t_{1/2}$ po	Chemprop MT	0.28	2.30
V_{ss}	Chemprop ST	0.56	0.27

The Alchemite method was also applied to the non-aggregated dataset where each compound had several replicate values of the PK parameter. The results are shown in Figure S2. 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 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 by the Alchemite(replace) method, a multi-task deep neural network with 2D chemical descriptors, where missing *in vitro* features were replaced with *in silico* values. Chemprop single-task model (Chemprop

ST) produced equivalent results (see Table 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 logit-transformed space). The RMSE of the model is approaching level of experimental error. The scatter plot of predicted versus observed values for logit-transformed F is shown in Figure 2. 65% and 84% of compounds are predicted within 2- and 3-fold error, respectively. Hence we can conclude that… (try to give key conclusion at end of paragraphs/sections)

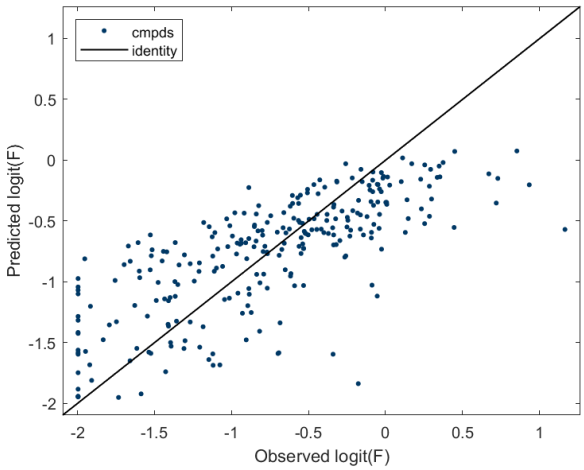

Figure 2. Predicted versus observed values for logit(F) on the test set for predictions made by Alchemite (replace) model. The identity line is solid black line.

3.1.3. Clearance model. Clearance is one of the most challenging parameters to optimise 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 by graph convolutions neural network method Chemprop applied in a single task setting (Chemprop ST), with Chemprop multi-task model (Chemprop MT) producing equivalent results (see Table 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 log-transformed space). The scatter plot of predicted versus observed values for log-transformed CL is shown in Figure 3(A). 78% and 94% of compounds are predicted within 2- and 3-fold error, respectively.

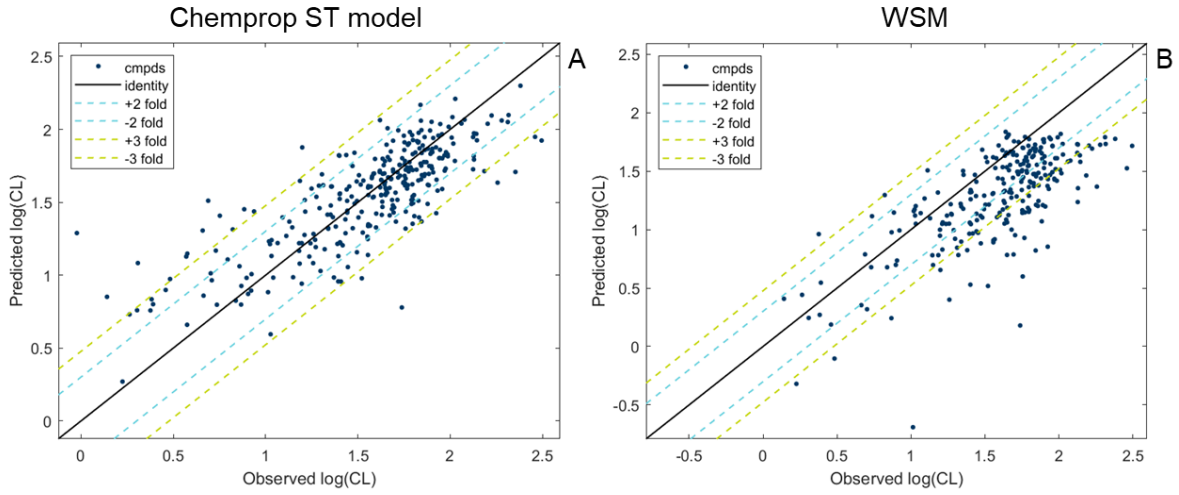

Figure 3: Predicted versus observed values for log(CL) on the test set for (A) predictions made by Chemprop ST model, (B) predictions made by WSM (259 compound subset of the test set). The identity line is solid black line, ± $log_{10}(2)$ lines corresponding to 2-fold error are dashed blue lines, ±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

3.1.4. Comparison with well-stirred model. The well stirred model (WSM) is a standard tool for *in vitro* and *in vivo* extrapolation (IVIVE) in drug discovery and is routinely applied in decision making for compound prioritisation and progression for *in vivo* testing and also to achieve understanding of the mechanism of clearance.[71-73] The WSM predicts clearance due to hepatic elimination, although the prediction is often assumed an approximation for the total clearance. The CL model was benchmarked against the WSM on the test set of 312 compounds. The predicted versus observed log-transformed CL values are shown in Figure 3 for both models. Predictions of the WSM are restricted by the rat liver blood flow (Qh=72 ml/min/kg or $\log_{10}$(Qh)=1.86), therefore the WSM predictions were available only for 259 compounds of the test set. As seen from Figure 3(B) 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 early discovery to guide compound prioritisation 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 WSM, 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 design, before compounds are synthesized and when ADME *in vitro* properties are not available, the performance of 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 optimisation 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, 'old' *in silico* ADME predictions were used instead of *in vitro* measurements. 'Old' ADME models were built before the test set compounds were synthesized. 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 model input. If the 'old' *in silico* predictions are used, there is an increase in RMSE between 5-30% depending on the PK parameter. E.g. for CL endpoint, RMSE = 0.35 for 'old' *in silico* predictions in comparison with RMSE = 0.27 for *in vitro* ADME values; for bioavailability F, RMSE = 0.57 and 0.47 for 'old' *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 'old' *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 input. This is of much practical relevance, since now compound PK in rat can be predicted solely based on chemical structure, without the necessity for experimental measurements.

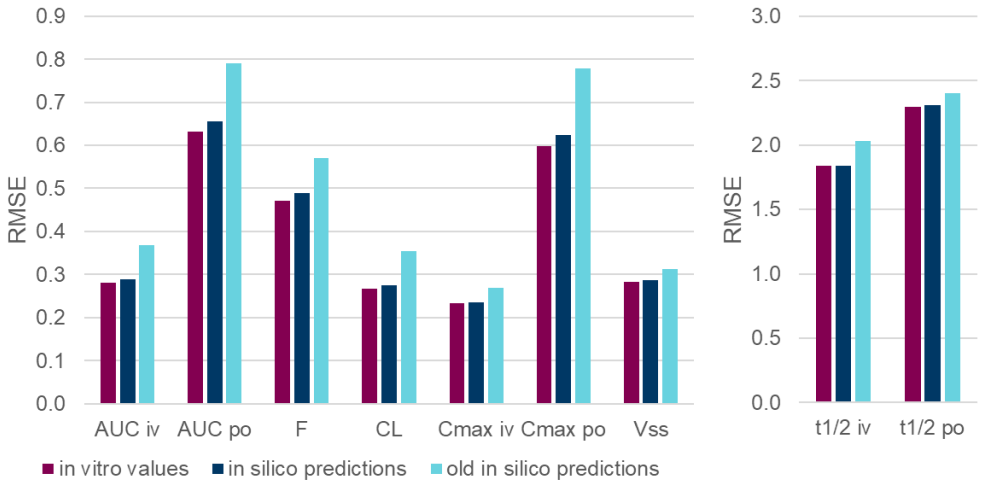

Figure 4: Performance of the Chemprop MT model on the test set utilising in vitro measurements for ADME features or corresponding in silico predictions. RMSE on the test set is shown when in vitro values (plum bars), in silico predictions (navy bars) and old in silico predictions (light blue bars) are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

3.1.6. Confidence in predictions. A good machine learning model provides an estimation of uncertainty in predictions as well as accurate predictions. The uncertainty quantification can enable detection of out-of-domain examples and identification of less reliable predictions. In this work, the explored algorithms that offer three different approaches to estimation of uncertainty. In the first approach, variability in prediction is captured by generating an ensemble of predictions. This approach is utilised by both deep neural network methods, Alchemite and Chemprop, as well as by GBoost, decision trees ensemble method. The second approach is inherent in Gaussian Processes, a Bayesian algorithm which is known to provide a useful quantification of uncertainty.[62,77] The output of the Gaussian Processes is not only a single point prediction but 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 conformal framework,[45,78] utilised in CPSign algorithm based on Support Vector Machine regression. These three approaches for uncertainty quantification were compared on the example of CL endpoint. The quality of different uncertainty estimates was evaluated using two metrics: ranking-based confidence curve with associated quantitative measure AUCO and calibration curve with associated quantitative measure AUCE . The confidence curves for four different CL models are shown in Figure 5. Clearly, all the four confidence curves are far from the perfect 'oracle' curve, that is the ranking order by predicted uncertainty does not correspond to ranking by real error of prediction. Chemprop MT method curve, shown in Figure 5(B), is closest to the 'oracle' curve and provides the best AUCO metric (AUCO=0.145). Both NN 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 NN methods, the uncertainty in prediction provided by the model significantly underestimated 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 Methods, section 2.4). Both Alchemite (replace) and Chemprop MT benefit from addition of the experimental uncertainty as shown in Figure 6(A) and Figure 6(B), respectively. GPR and CPSign models, on the other hand, produce close to perfect calibration curves, Figure 6(C-D). For GPR and CPSign, the addition of the experimental uncertainty was not needed, the model uncertainty estimation incorporates all sources of uncertainty and represents the total uncertainty. GPR technique estimates uncertainty using

Bayesian approach, CPSign involves empirical estimation via conformal prediction framework. The best calibration curve is provided by GPR model with AUCE = 0.026.

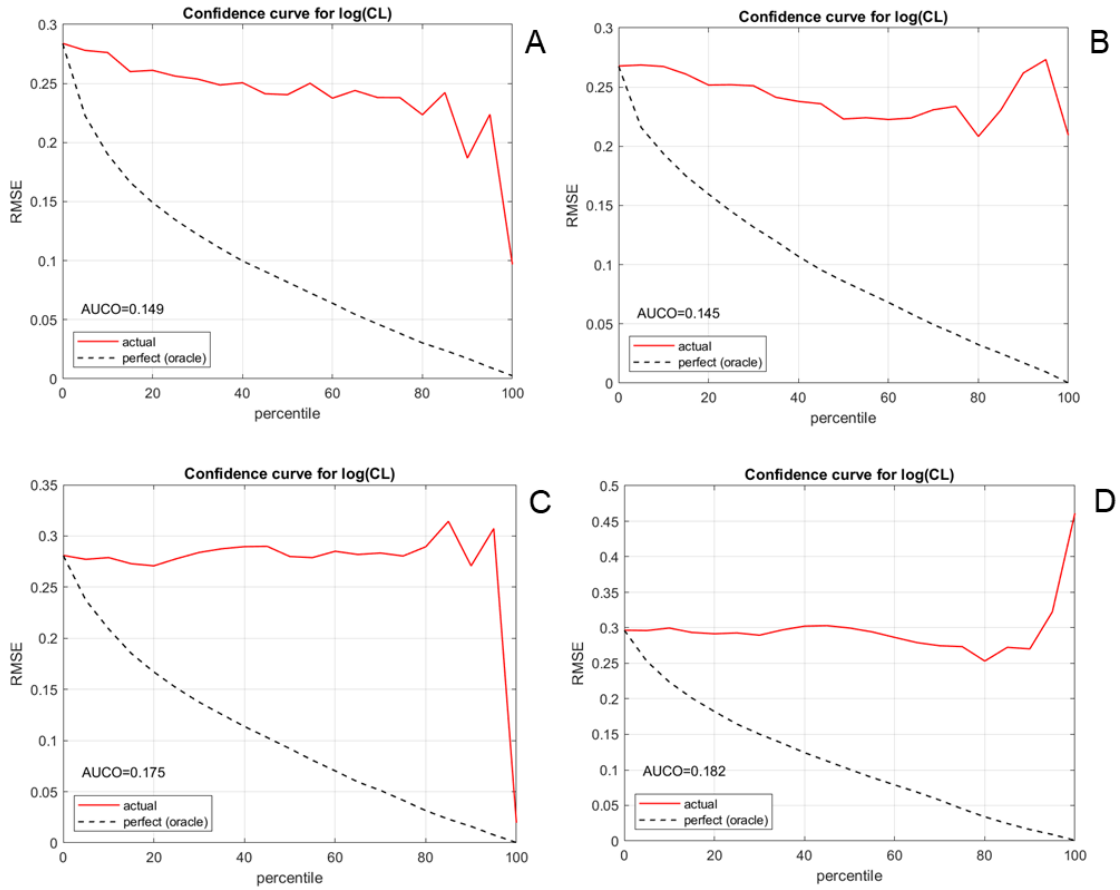

Figure 5: The confidence curves and corresponding AUCO values for CL endpoint obtained on the test set using predictions and uncertainty estimates from different models - (A) Alchemite (replace), (B) Chemprop MT, (C) GPR, (D) CPSign. The oracle curve is a dashed black line.

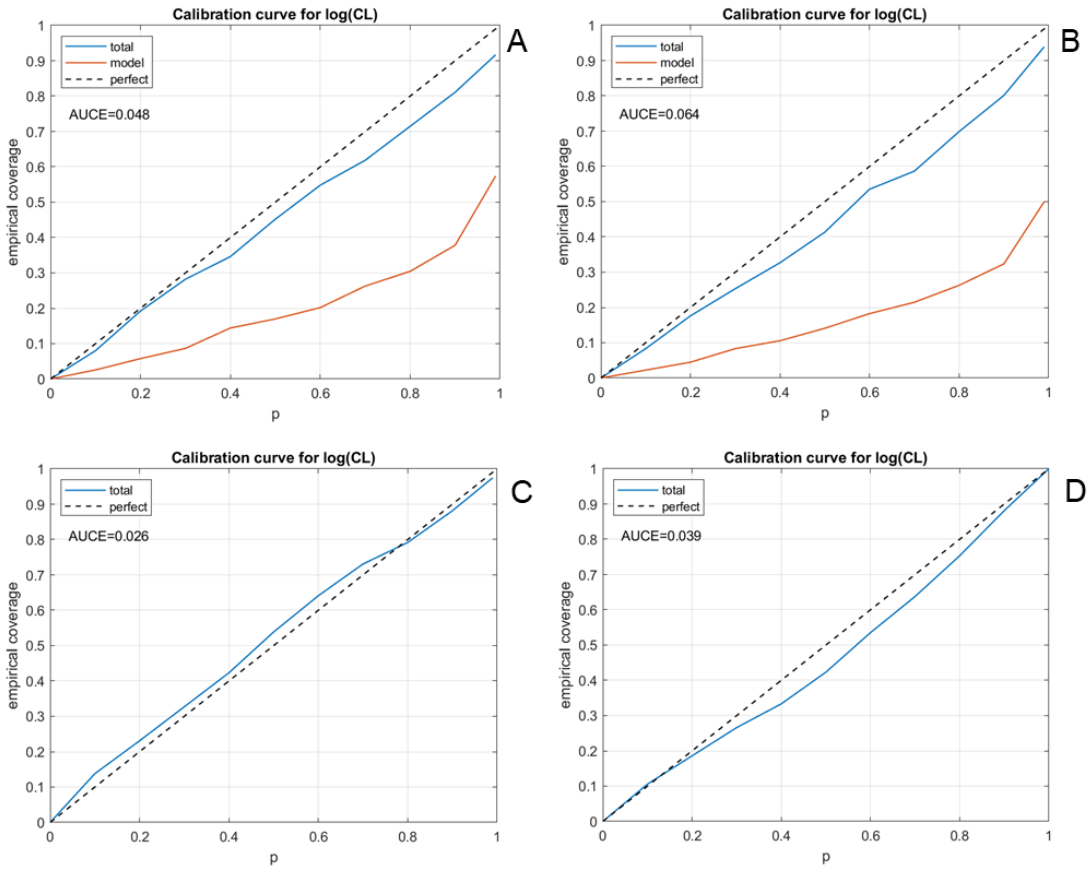

Figure 6: The calibration curves and corresponding AUCE values for CL endpoint obtained on the test set using predictions and uncertainty estimates from different models - (A) Alchemite (replace), (B) Chemprop MT, (C) Gaussian processes, (D) CPSign. The confidence curves based on the model uncertainty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

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-times are shown in Figure 7, summarising the performance of the model on all the test set compounds (312 compounds for iv; 279 for po). 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 measurement tolerance, 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 median R^2 of -0.78 (median RMSE 0.54 log units). This is likely to be because po PK is more complex than iv, 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 modelling task. We therefore progress our analysis only of the iv dosing curves.

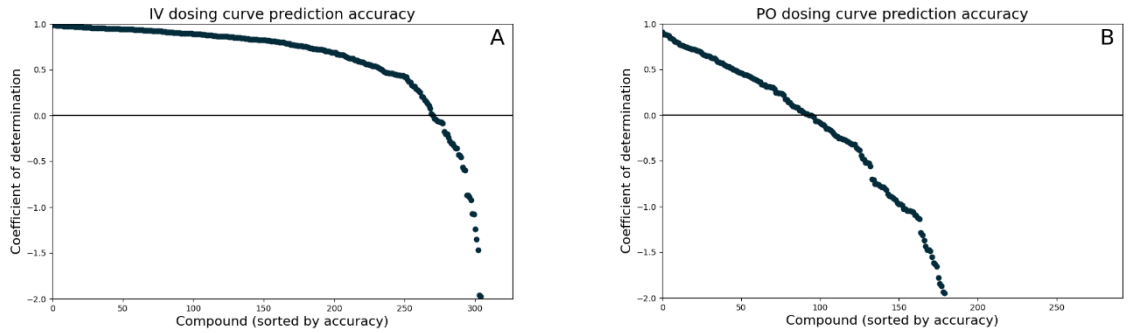

Figure 7: Profiles of accuracy in prediction of PK curves, with R^2 calculated for time curves for which both experimental and modelled values are available, averaged across replicates, for both iv dosing (A) and po dosing (B). Profiles are truncated at R^2=-2.

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.

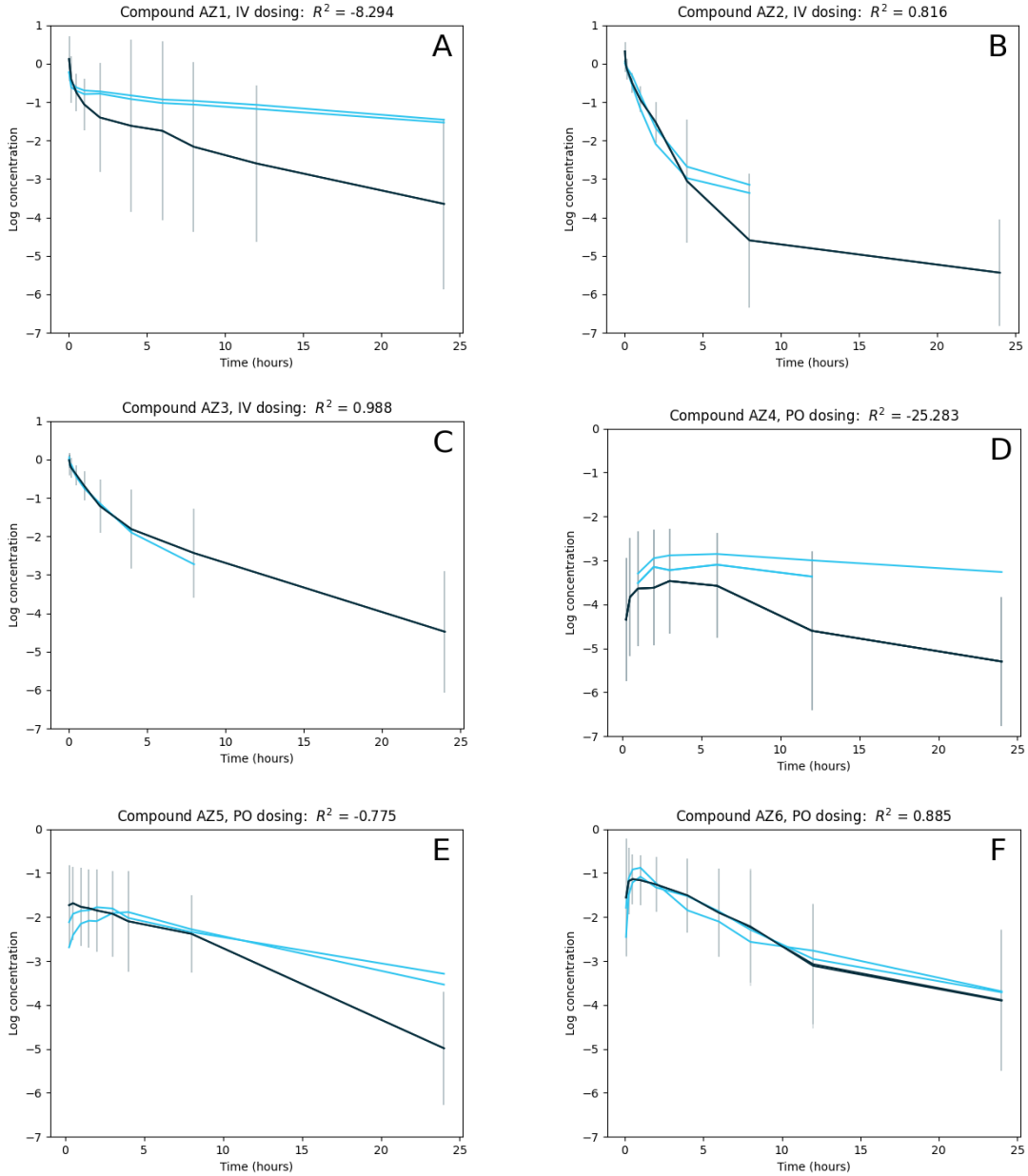

Figure 8: A selection of iv and po dosing curves: experimental data is shown in light blue, including multiple replicates per compound, and the predicted curves are shown in dark blue, with uncertainty in prediction shown by the vertical grey lines. Coefficient of determination measures for the accuracy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

3.2.2. *Calculation of parameters from* curves. To enable comparison with the results in Section 3.1, PK parameters were generated from the predicted curves and compared to the (experimental) PK parameters used for modelling in Section 3.1. These PK parameters had been generated from the true experimental data using a semi-manual process involving cleaning of the underlying data: however for all PK parameters except V_{ss} the Pearson correlation between the semi-manual generation and the fully-automated generation using MATLAB exceeded 0.97, indicating the semi-manual process made only small differences to the PK parameter generation. The results for the iv curves are summarised 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 predicted 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 late time behaviour of the curve, which, as discussed above, is less accurately captured by the model than the earlier time behaviour, 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 from chemical structure representations and experimentally measured ADME properties. We also performed evaluation of multiple machine leaning algorithms and approaches to missing data imputation.

The models are based on a dataset of over 3,000 diverse compounds from multiple drug discovery projects for various therapeutic applications, measured in the same lab using single assay post-intravenous and oral dosing. The input experimental features of the models include ADME and physicochemical properties describing compounds lipophilicity, solubility, permeability, intrinsic metabolic clearance, and plasma protein and hepatocyte binding. For the chemical structure representation we explored graph convolutional neural networks that encode molecule chemical graph structure, 1D and 2D molecular property descriptors and signature descriptors. We applied state-of the-art AI approaches, such as graph convolutional neural network Chemprop and deep learning technology Alchemite, as well as traditional machine learning algorithms such as Gaussian Processes, Support Vector Machines and Gradient Boosting Tree ensembles. Because some of the experimental ADME data is missing in the dataset and to allow for such situations for future predictions, we investigated two data imputation approaches – the Alchemite algorithm and the 'replace' approach (utilisation of *in silico* predictions for ADME properties generated by internal global models in the absence of experimental data). 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 dataset. 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 good 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 optimise 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 prioritisation. On the test set, the model, predicting total *in vivo* clearance, achieved much higher accuracy with R^2=0.63 versus R^2=-0.11 for the WSM however it should be noted that WSM only estimates hepatic metabolic clearance. Therefore, the CL model provides an accurate and useful tool for decision making in early discovery, and being able to predict values higher than the liver blood flow, it complements current DMPK tools used for PK prioritisation. 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, good accuracy 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 can be used at the point of design, before compounds are synthesized and before ADME *in vitro* properties become available; we observed relatively small decrease in the accuracy.

As well as 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 intravenous dosing is good (the median of individual curve R^2 equals 0.82), but the prediction of curves with oral dosing is poor, perhaps due to higher variability in oral dosing curves data. PK parameters estimated from predicted intravenous curves are slightly less accurate overall than those predicted by the models directly.

The models provide a powerful way to guide the design of molecules with optimal PK profiles, to enable the prediction of virtual compounds, and to drive prioritisation of compounds for *in vivo* assays. Furthermore, the developed AI approach is a stepping stone for the prediction of human PK, ultimately leading to the design of molecules with a desired multi-objective profile early in drug discovery, which will increase efficiency and reduce compound attrition.

ABBREVIATIONS

REFERENCES

1. Ruiz-Garcia, A. *et al.* Pharmacokinetics in Drug Discovery. *J. Pharm. Sci.* **2008** (97) 654-690.
2. Sturm, N. *et al.* Industry-scale application and evaluation of deep learning for drug target prediction. *J. Cheminform.* **2020** (12) 26.
3. Mervin, L.H. *et al.* Target prediction utilising negative bioactivity data covering large chemical space. *J. Cheminform.* **2015** (7) 51.
4. Blaschke, T. *et al.* REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.* **2020** (in press, DOI: 10.1021/acs.jcim.0c00915).
5. Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery – What is Realistic, What are Illusions? Part 1: Ways to impact, and why we are not there yet. *Drug Discov. Today* **2021** (in press, DOI xxx).

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 used for AI in drug discovery. *Drug Discov. Today* **2021** (in press, DOI …).
7. Hughes, J. P. *et al.* Principles of early drug discovery. *Br. J. Pharmacol.* **2011** (162) 1239–1249.
8. Chen, H. *et al.* The rise of deep learning in drug discovery. *Drug Discov. Today* **2018** (23) 1241-1250.
9. Paul S.M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010** (9) 203-214.
10. Scannell, J.W.; Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLoS One* **2016 (**11) e0147215.
11. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov* **2020** (19) 353–364.
12. Gillette, J. R. Factors affecting drug metabolism. *Annal. N. Y. Acad. Sci.* **1971** (179) 43 – 66.
13. Pang, K.S. *et al.* 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.
14. Sager, J.E. *et al.* 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.
15. Davies, M. *et al.* Improving the Accuracy of Predicted Human Pharmacokinetics: Lessons Learned from the AstraZeneca Drug Pipeline Over Two Decades. *Trends Pharm. Sci.* **2020** (41) 390 – 408.
16. Morgan, P. *et al.* Impact of a five-dimensional framework on R&D productivity at AstraZeneca *Nat. Rev. Drug Discov.* **2018** (17) 167-181.
17. Williamson, B. *et al.* Evaluation of the disconnect between hepatocyte and microsome intrinsic clearance and in vitro in vivo extrapolation. *Drug Metab. Disp.* **2020** (48) 1137-1146.
18. Lowe, Jr. E.W. *et al.* Comparative Analysis of Machine Learning Techniques for the Prediction of the DMPK Parameters Intrinsic Clearance and Plasma Protein Binding. *Proc. 4th Int. Conf. Bioinf. Comp. Biol.* **2012**
19. Wang, Y. *et al.* In Silico Prediction of Human Intravenous Pharmacokinetic Parameters with Improved Accuracy. *J. Chem. Inf. Model.* **2019** (59) 3968−3980.
20. Lombardo, F. *et al.* Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug Metab. Dispos.* **2018** (46) 1466-1477.
21. Lombardo, F. *et al.* 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*. **2020** (110) 500 – 509.
22. Kamiya, Y. *et al.* Physiologically Based Pharmacokinetic Models Predicting Renal and Hepatic Concentrations of Industrial Chemicals after Virtual Oral Doses in Rat. *Chem. Res. Toxicol.* **2020** (33) 1736–1751.
23. Bassan, A. *et al.* In silico approaches in organ toxicity hazard assessment: 1 current status and future needs in predicting liver toxicity. *Regul. Pharm. Tox.* **2021** (submitted)
24. Bassan, A. *et al.* In silico approaches in organ toxicity hazard assessment: current status and future needs for predicting heart, kidney and lung toxicities. *Regul. Pharm. Tox.* **2021** (submitted)
25. Schneckener, S. *et al.* 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.

26. Feinberg, E. N. *et al*. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020** (63) 8835–8848.
27. Irwin, B. W. J., Julian Levell, Thomas M Whitehead, Matthew D Segall, Gareth J Conduit Practical Applications of Deep Learning To Impute Heterogeneous Drug Discovery Data. *J. Chem. Inf. Model.* **2020** (60) 2848–2857.
28. Ye, Z. *et al.* An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol. Pharm.* **2019** (16) 533-541.
29. 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.
30. Kosugi Y. and Hosea N. Prediction of Oral Pharmacokinetics Using a Combination of In Silico Descriptors and In Vitro ADME Properties. *Mol. Pharm.* **2021** (18) 1071−1079.
31. Yang K *et al*. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019** (59) 8, 3370-3388.
32. Winiwarter S *et al*. Time dependent analysis of assay comparability - a novel approach to understand intra- and inter- site variability over time. *J Computer-Aided Mol. Des.* **2015** 29 (9), 795-807. DOI: 10.1007/s10822-015-9836-5
33. Wenlock MC, Potter T, Barton P, Austin RP. A method for measuring the lipophilicity of compounds in mixtures of 10. *J Biomol Screen*, **2011** (16), 348–355.
34. Wan H, Holmén AG. High throughput screening of physiochemical properties and in vitro ADME profiling in drug discovery. *Comb Chem High Throughput Screen*, **2009** (12), 315–329.
35. 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, 5) 1601–1609; DOI: 10.1021/acs.molpharmaceut.6b01059
36. A.-K. Sohlenius-Sternbeck, L. Afzelius, P. Prusis, J. Neelissen, J. Hogstraate, J. Johansson, E. Floby, A. Bengtsson, O. Gissberg, J. Sternbeck, C. Petersson. Evaluation of the human prediction of clearance from hepatocytes and microsome intrinsic clearance for 52 drug compounds. *Xenobiotica* **2010** (40) 637-649.
37. Wenlock MC, Carlsson LA. How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. *J. Chem. Inf. Model.* **2015** (55) 125-134.
38. D.G. Temesi, S. Martin, R. Smith, C. Jones, B. Middleton. High-throughput metabolic stability studies in drug discovery by orthogonal acceleration time-of-flight (OATOF) with analogue-to-digital signal capture (ADC). *Rapid Communications in Mass Spectrometry* **2010** (24) 1730-1736.
39. H. Wan, F. Bergström. High Throughput Screening of Drug Protein Binding in Drug Discovery. *Journal of Liquid Chromatography & Related Techniques* **2007** (30) 681-700.
40. N.J. Waters, R. Jones, G. Williams, B. Sohal. Validation of a Rapid Equilibrium Dialysis Approach for the Measurement of Plasma Protein Binding. *Journal of Pharmaceutical Sciences* **2008** (97) 4586.
41. R.P. Austin, P. Barton, S. Mohmed, R.J. Riley. The binding of drugs to hepatocytes and its relationship to physicochemical properties. *Drug Metabolism and Disposition* **2005** (33) 419-425.
42. Olsson T and Sherbukhin V. Synthesis and structure administration (SaSA). AstraZeneca R&D Mölndal. **2002**.
43. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot,

Édouard Duchesnay. Scikit-learn: Machine Learning in Python, *JMLR* **2011** (12), pp. 2825-2830.
44. Faulon J-L, Visco DP, and Pophale RS. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *Journal of Chemical Information and Computer Sciences* **2003** (43) 707–720.
45. Norinder U, Carlsson L, Boyer S, and 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.
46. CPSign Documentation. Available online at: https://arosbio.com/cpsign/docs/latest/# (Accessed 2020-12-08)
47. Alvarsson J, Eklund M, Andersson C, Carlsson L, Spjuth O, Wikberg JES. Benchmarking study of parameter variation when using signature fingerprints together with support vector machines. *J. Chem. Inf. Model.* **2014** (54) 3211–3217.
48. Oprisiu I, Winiwarter S. In silico ADME modeling. *Systems Medicine: Integrative, Qualitative and Computational Approaches.* **2021** (2) 208-222.
49. 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, 10.1021/acs.jcim.8b00768
50. Kier LB and Hall LH (**1986**) Molecular connectivity in structure-activity analysis. New York: Wiley.
51. ACD/Labs software, 2015, Advanced Chemistry Development, Inc., Toronto, Ontario, Canada.
52. ClogP, 4.3, Pomona College and BioByte, Inc., Claremont, CA, US.
53. P. Bruneau. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *Journal of Chemical Information and Computer Sciences*. **2001** (41) 1605-1616.
54. 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 (11−12), 960−72
55. Duvenaud, David K., et al. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*. **2015**.
56. Kearnes, Steven, et al. Molecular graph convolutions: moving beyond fingerprints. *J Computer-Aided Mol. Des.* **2016** (30.8) 595-608.
57. StarDrop StarDrop v?, Optibrium Ltd., Cambridge, UK
58. Daylight SMARTS
59. Landrum, G. RDKit: Open-Source Cheminformatics; **2006**. https://rdkit.org/docs/index.htm
60. Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on Machine Learning. **2010**, 807– 814.
61. MacKay, D. J. C. Information Theory, Inference, and Learning Algorithms; Cambridge University Press: Cambridge, United Kingdom, 2003.
62. Rasmussen, C. E.; Williams, C. K. I. Gaussian Processes for Machine Learning; The MIT Press: Cambridge, MA, 2006.
63. Obrezanova O, Csanyi G, Gola JM, Segall MD. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J Chem Inf Model.* **2007** 47(5):1847-57. doi: 10.1021/ci7000633.
64. Obrezanova O, Gola JM, Champness EJ, Segall MD. Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J Comput Aided Mol Des.* **2008** 22(6-7):431-40. doi: 10.1007/s10822-008-9193-8.
65. Burden, F. R. Quantitative Structure-Activity Relashionship Studies Using Gaussian Processes. J. Chem. Inf. Comput. Sci. 2001, 41, 830-835.

66. MATLAB: 2019a, Natick, Massachusetts, The Mathworks, Inc., MATLAB version R2019a, 2019. https://www.mathworks.com/
67. Conduit, B. D., Jones, N. G., Stone, H. J., Conduit, G. J., Design of a nickel-base superalloy using a neural network. *Materials and Design* **2017**, 131, 358, 10.1016/j.matdes.2017.06.007
68. Bergstra, J., Yamins, D., Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. TProc. of the 30th International Conference on Machine Learning (ICML 2013), June **2013**, pp. I-115 to I-23.
69. Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J Chem Inf Model.* **2020** 60 (6), 2697-2717. DOI: 10.1021/acs.jcim.9b00975
70. 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(8), 1246-1252. http://dx.doi.org/10.1124/dmd.116.070763 PMID: 27084892
71. Rowland
72. Williamson B, Colclough N, Fretland AJ, Jones BC, Jones RDO, McGinnity DF. Further Considerations Towards an Effective and Efficient Oncology Drug Discovery DMPK Strategy. *Curr Drug Metab.* **2020**;21(2):145-162. doi: 10.2174/1389200221666200312104837.
73. Riley RJ, Mcginnity DF & Austin RP. A unified model for predicting human hepatic, metabolic clearance from in vitro intrinsic clearance data in hepatocytes and microsomes. *Drug Metab Dispos* **2005** (33) 1304-11.
74. Noncompartmental analysis. https://uk.mathworks.com/help/simbio/ug/non-compartmental-analysis.html
75. MATLAB R2019a, The Mathworks, Inc., Natick, Massachusetts, MATLAB version 9.6.0.1072779 (R2019a), 2019.
76. Yang J, Jamei M, Yeo KR, Rostami-Hodjegan A & Tucker GT. Misuse of the well-stirred model of hepatic drug clearance. *Drug Metab Dispos* **2007** (35) 501-2.
77. Brian Hie, Bryan D. Bryson, Bonnie Berger. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design, *Cell Systems* **2020** (11), Issue 5, 461-477; https://doi.org/10.1016/j.cels.2020.09.007.
78. Isidro Cortés-Ciriano and Andreas Bender. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2019** (59) 1269−1281.