Actionable predictions of human pharmacokinetics at the drug design stage

Leonid Komissarov,*† Nenad Manevski,† Katrin Groebke Zbinden,† Torsten Schindler,† Marinka Zitnik,‡ and Lisa Sach-Peltason*†

†Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, 4070 Basel, Switzerland
‡Department of Biomedical Informatics, Harvard University

E-mail: leonid.komissarov@roche.com; lisa.sach-peltason@roche.com

Abstract

We present a novel computational approach for predicting human pharmacokinetics (PK) that addresses the challenges of early-stage drug design. Our study introduces and describes a large-scale dataset of 11 clinical PK endpoints, encompassing over 2700 unique chemical structures to train machine learning models. To that end multiple advanced training strategies are compared, including the integration of in vitro data and a novel self-supervised pre-training task. In addition to the predictions, our final model provides meaningful epistemic uncertainties for every data point. This allows us to successfully identify regions of exceptional predictive performance, with an Absolute Average Fold Error (AAFE/GMFE) of less than 2.5 across multiple endpoints. These advancements represent a significant leap towards actionable PK predictions, which can be utilized early on in the drug design process to expedite development and reduce reliance on nonclinical studies.

Introduction

Human pharmacokinetics (PK) is an integral part of drug efficacy and safety. Therefore, successful computational predictions of PK would benefit drug discovery efforts by increasing the quality and speed of drug design and reducing the number of nonclinical PK studies. Human PK is governed by passive and active disposition processes. Passive processes, driven by laws of thermodynamics, include dissolution and diffusion in biological fluids, permeability across biological membranes, and binding to tissues and organs. Active processes, characterized by the use of energy such as ATP-hydrolysis, include activities of drug-metabolizing enzymes and transporters, as well as physiological processes like gut motility and blood flow. Conceptually, one could theorize that most of these drug disposition processes are encoded by the chemical structure of the compound, either through global molecular properties (e.g. logD and pKa) or specific interactions with active sites of biomolecules. Remaining disposition processes like blood and lymph flows can be considered physiological constants for a given species, even if interindividual differences exist. Therefore, given a sufficiently large and chemically diverse dataset of historical PK studies, one could attempt to predict PK of novel compounds based on chemical structure alone.
Several recent studies have confirmed this hypothesis — that PK properties can be predicted based on chemical structure — using both clinical\(^1\text{–}^3\) and nonclinical\(^4,^5\) PK datasets. Moreover, the predictive performance of these computational models appears to be in the range of classical methods, such as in vitro-in vivo extrapolation (IVIVE) and allometric scaling. For example Miljković et al. report high-performance predictive models for human AUC\(_{PO}\), C\(_{max}\), and V\(_{ss}\), outperforming classical predictions.\(^1\) More recently Lombardo et al.\(^6\) have demonstrated the same for the prediction of clearance. Kosugi and Hosea,\(^7\) Keefer,\(^8\) and Andrews-Morger et al.\(^9\) found that computational predictions of CL\(_p\) improve upon classical IVIVE. Similarly, our previous work on computational predictions of mouse and rat PK studies showed that most properties can be predicted within 2–3-fold error,\(^4\) well within the range of best classical methods.\(^10,^11\) Learning from PK history offers an additional advantage of capturing complex disposition phenomena, such as paracellular and lymphatic absorption, extrahepatic metabolism, and transporter-mediated disposition, often understood only at later stages of drug development.

Despite the promise of emerging computational methods, it is fair to acknowledge their current limitations. Firstly, existing human PK datasets are small (<1400 compounds), focused on a single property (e.g. only CL\(_p\) or V\(_{ss}\)), and slow-growing. This is in stark contrast to the practically unlimited chemical space of drug discovery. Secondly, computational models predicting a single PK property ignore much of the information landscape, such as correlated PK properties, similar PK results from other species, and corresponding in vitro data. Thirdly, model performance metrics focus on average prediction quality but may not represent model performance for a given compound of interest. These challenges, taken together, limit the practical usefulness of such models, effectively confining them to experimental scenarios.

By focusing on the aforementioned challenges, the present study aims to advance computational predictions of human PK. Our goal is to provide actionable models that could be deployed in drug discovery, especially focusing on the drug design stage, where only chemical structure of the compound is available. For this purpose we curated one of the largest human PK datasets extracted from approval packages and published studies (46934 entries and 2702 unique chemical structures). Potential for transfer learning was investigated from several angles, including multitask predictions of all key PK properties, integration with in vitro and pre-clinical in vivo data, as well as supervised and self-supervised pre-training. Significant effort was invested in methods for quantifying model (i.e. epistemic) uncertainty, aiming to provide well-calibrated models that correlate prediction uncertainty with observed prediction error.

**Methods**

**Dataset curation & analysis**

Sources for the clinical PK dataset were Elsevier’s Pharmapendium (PP) and Clarivate Cortellis Drug Discovery Intelligence (CDDI). Both platforms offer data extracted from regulatory bodies’ approval packages (U.S. Food and Drug Administration and European Medicines Agency). In addition, CDDI includes data from published studies that are not part of any official approval package. Raw data snapshots from both providers were obtained in August 2022 and March 2023 for PP and CDDI, with initial records containing 1.2m and 1.6m pharmacokinetic entries, respectively. Rigorous filtering was performed to distill a machine learning-ready data set from the initial data. We restrict the entries to

0. Initial data (25625)
1. Humans (5190)
2. Healthy adult population (3864)
3. Single-compound and single-dose regimens (3214)
4. Plasma, serum or blood compartments (3203)
5. *Per os* (PO) and intravenous (IV) routes of administration (3022)
6. Dose standardization (2931)
7. Molar dose < 0.01 mol (corresponding to a dose of 5 g for a drug with a molar weight of 500 g/mol) (2874)
8. SMILES standardization (2718)
9. Molar weight ≤ 1000 g/mol (2624)
10. Filter for relevant PK parameters & standardization (2499)
11. Outlier removal (2495)

At each step the remaining number of unique chemical structures in the CDDI data set is shown in parentheses (relative attrition for Pharmapendium is similar and therefore omitted). For PP, we additionally remove all radiolabeled studies and studies with concomitants that could influence the pharmacokinetic profile of a compound. We limit ourselves to the following pharmacokinetic parameters. *Per os*: Area under the concentration-time curve (AUC<sub>PO</sub> [mol h/L]), bioavailability (F [%/100]), apparent clearance (CL/F [L/h]), apparent volume (V/F [L]), maximum concentration (C<sub>max</sub> [mol/L]), time at maximum concentration (T<sub>max</sub> [h]), elimination half time (T<sub>half,PO</sub> [h]). *Intravenous*: Area under the concentration-time curve (AUC<sub>IV</sub> [mol h/L]), clearance (CL [L/h]), volume (V [L]), elimination half time (T<sub>half,IV</sub> [h]). We found that annotations of the observed volume in the original data were often inaccurate. For this reason no distinction between volume of distribution and volume of the central compartment is made. Parameters with units differing from the above are converted, assuming a standard human body weight of 70 kg where applicable. Note that AUC and C<sub>max</sub> values are not dose-normalized, as dose is used as an additional input parameter for our models. All chemical structures are standardized with the ChEMBL structure pipeline, applying the standardization as well as parent structure generation functions consecutively. Chemical structures that contain more than one fragment after standardization are discarded. We do not differentiate between active compounds and prodrugs. Data entries can contain metabolite structures: when an administered and measured structures differ, the latter is considered. Outliers in parameter values are removed as follows: First, we group all replicate measurements with the same standardized chemical structure, dose, route of administration and PK parameter. Within each group the median absolute deviation-adjusted z-score<sup>13</sup> is applied to detect outliers. We use a z-score threshold of 2 (in analogy to filtering all entries that are outside of two standard deviations for normally-distributed data). For each group, the median of all remaining values is used as the ground truth value. Second, after log<sub>10</sub>-transformation (logit for bioavailability) of all ground truth values, global outliers are removed by applying an elliptic envelope<sup>14</sup> with a contamination of 1% (as implemented in the scikit-learn package<sup>15</sup> v1.2.2) to each parameter distribution individually. The resulting data set contains a total of 2702 unique chemical structures and 46934 entries, where one entry comprises the chemical structure, dose, route of administration and pk parameter.

**Estimation of experimental variability**

Before outlier removal, we estimate the experimental variability in each parameter as follows to provide a lower bound for the accuracy of our models. As mentioned above, replicate measurements are identified by grouping parameter values by the same chemical structure, dose, PK parameter and route of administration. Whenever a group contains six or more measurements (as a trade-off between sample size and coverage; covering 990 out of the total 2700 compounds), the experimental ground truth is approximated as the median of all measurements within the group. We can then calculate any suitable accuracy metric on the experimental ground truth versus the individual values and thus use it to approximate the experimental error. Here we use the absolute average fold error (AAFE) (see Model performance metrics Section for definition).
**Chemical space analysis**

Because our combined data set originates from two sources, holding \( n \) and \( m \) chemical structures, respectively, we were interested in a quantitative measure of similarity between the two sets. To tackle this question, we propose adapting a solution to the optimal transport problem.\(^{16}\) To the best of our knowledge, this is the first application of the optimal transport solution to the quantification of similarity for two sets of chemical structures. Conceptually, our proposed solution differs from parametric approaches like the k-nearest neighbor algorithm in that it does not require an *a priori* specification of the number of neighbors to be considered. Instead, the number of assignments between points is variable as the algorithm aims to reduce the total cost associated with moving one distribution onto the other. First, we solve for the optimal transport matrix \( T^* \)

\[
T^* = \arg \min_{T \in \mathbb{R}^{n \times m}} \sum_{i,j} T_{i,j} M_{i,j}
\]

subject to the constraints \( \sum_{i} T_{i} = a \in \mathbb{R}^n, \sum_{j} T_{j} = b \in \mathbb{R}^m, \sum_{i,j} T_{i,j} = 1 \), where \( M \) is a pairwise distance matrix between all elements in the two sets of chemical structures, and \( a \) and \( b \) are weights vectors associated with each element in either set. In practice, we use the Python Optimal Transport (POT, v0.9.1) package\(^{17}\) to solve the above equation and calculate the distance matrix in the Tanimoto/Jaccard\(^{18,19}\) distance space, assuming a uniform distribution of weights, i.e. \( a_i = \frac{1}{n} \) and \( b_i = \frac{1}{m} \). Chemical structures are encoded as Feature Morgan fingerprints with a radius of 2, folded to 1024 bits (hereinafter FCFP4), as implemented in RDKit\(^ {20}\) (v2023.09.3). We then use the earth mover’s distance,\(^{21}\) defined as

\[
EMD = \sum T_{i,j}^* M_{i,j}
\]

as a distance measure for the two sets. A larger \( EMD \) corresponds to a higher dissimilarity of two sets of chemical structures and vice versa. Analogous to the Tanimoto similarity of two molecules, the \( EMD \) remains bounded to the range of \( EMD \in [0, 1] \) in this application, conveniently allowing for the same definition of a set similarity coefficient \( S \) as

\[
S = 1 - EMD
\]

In addition to the quantitative similarity measure, we provide a visual representation of the chemical space by performing dimensionality reduction on the FCFP4 fingerprints, embedding them into a two-dimensional space with the Uniform Manifold Approximation and Projection\(^ {22}\) (UMAP) algorithm. Aiming to capture a large, yet relevant chemical space, we use a random selection of 250 thousand chemical structures from the ChEMBL database\(^ {23,24}\) as training data for the embedding. Following the fitting procedure, chemical structures from CDDI and PP are embedded into the same space inductively. We use the Tanimoto/Jaccard distance and set \( n_{\text{neighbors}}=150 \) and \( \text{min}_\text{dist}=0 \) for the fitting process.

**Machine learning models**

We assume two input modalities for all machine learning (ML) models: Chemical structure, initially encoded as SMILES strings, and molar dose. Performance of all models is assessed using k-fold cross-validation (k=5) on 90% of the data set (random selection based on the SMILES string). The remaining fraction is reserved for testing. We experiment with four different splitting strategies for cross-validation, each based on the chemical structure: (1) random split, (2) Murcko scaffold\(^{25}\) split, (3) generic Murcko scaffold split and (4) SCINS\(^ {26}\) split. In the case of generic Murcko scaffolds, all atoms are converted to carbon, and all bonds are converted to single bonds. Intuitively, given a set of compounds that homogeneously cover an arbitrary chemical space, we expect the splitting strategies to be increasingly more challenging moving from (1) to (4), as the number of unique scaffolds decreases.

Baseline predictive model performance is compared across three different frameworks: XGBoost\(^ {27}\) (v1.7.4), Chemprop\(^ {28,29}\)
(v1.5.2) and an in-house model based on the GraphGPS\textsuperscript{30} architecture (hereinafter GPS). We motivate our choice of the former two models based on their wide adoption to the task of chemical property prediction.\textsuperscript{31–35} GPS was selected as a more recent model that is among the top scoring models in the Large-Scale Challenge\textsuperscript{36} of the Open Graph Benchmark\textsuperscript{37} (specifically PCQM4Mv2). Conceptually, XGBoost applies gradient boosting to an ensemble of decision trees to arrive at a prediction. Input to the XGBoost models are FCFP4 fingerprints with the molar dose appended to each feature vector.

Chemprop and GPS are graph neural networks (GNNs) – a formalism that allows learning of relevant structural features by encoding and passing information along the nodes and edges of the molecular graph. For Chemprop, the built-in featurizer is used to convert SMILES strings into input vectors. Initial features for the GPS model are generated from a number of atom and bond properties – additional details are provided in Sec. S4.2 of the Supporting Information. We find empirically that the following two modifications to the original GPS recipe yield slightly improved performance: First, convolutions on the graph structure are implemented with the GATv2\textsuperscript{38} layer. Second, in addition to the node attributes, edge attributes are passed through a multi-layer perceptron at each GraphGPS layer (see S4.1 for more details). Unless explicitly stated otherwise, the GPS and Chemprop models are set up to predict labels in a multi-task fashion (i.e. one model is trained to predict all PK parameters simultaneously). In the case of XGBoost a separate model is trained for each PK parameter. Model performance is compared after hyperparameter optimization.

For Chemprop, the built-in hyperparameter optimization routine is used. The optuna package\textsuperscript{39} (v3.1.1) is used to optimize the GPS and XGBoost hyperparameters. Please for to the Supporting Information (Sec. S4.1) for hyperparameters and the architectural details of our final GPS model.

Leveraging additional data for training

The ability to handle sparse data as well as save, modify, and combine different sets of weights into a predictive model makes neural network-based formalisms especially appealing with regard to more elaborate modes of training. In addition to training on the target human PK data alone, we experiment with utilizing various larger data sets as to create a more accurate model across a wide chemical space. Throughout all experiments in this section we use the hyperparameter-optimized GPS model.

Pre-training is a popular technique in machine learning applications where the training data of interest (i.e. ‘downstream’) is too scarce to guarantee good predictive performance across a larger input space. The technique aims to alleviate this issue by leveraging one or multiple larger data sets (i.e. ‘upstream’) where the training objective is not the same, but similar to the downstream task. In doing so, one expects to steer the model parameters into a favorable space such that consecutive training on the downstream data results in better performance.\textsuperscript{40} This strategy has been successfully applied to the task of molecular property predictions.\textsuperscript{41–43} Pre-training can be further divided into supervised and self-supervised. The former utilizes upstream data that contains training labels, the latter does not require training labels \textit{a priori} and instead generates them from the input. Here we experiment with both approaches.

As an alternative to supervised pre-training, multiple labeled data sets can be combined for direct training of a multi-task model. This can be useful when the auxiliary labels (i.e. the ones that are not of primary interest) are related to the tasks of interest. One downside to this approach is the increased computational cost as models need to be trained from scratch on large data sets. In contrast, a pre-trained model can be stored and fine-tuned with less effort.

For datasets where labels are unavailable we implement and apply two self-supervised pre-training strategies to the GPS model (Fig. 2f). First, we adapt the MolCLR formalism.\textsuperscript{44} The method follows a contrastive learning approach
where a chemical graph is modified by atom and/or bond removal combined with feature masking. The model is then trained to recognize the unmodified-modified pair of graphs that belong to the same chemical structure. This aims to enforce that for two similar structures, the learned embeddings should be similar also. One drawback of contrastive learning approaches is that modifications to the graph structure of a compound usually result in chemically invalid structures. We therefore implement a strategy based on the Breaking of Retrosynthetically Interesting Chemical Substructures (BRICS) algorithm. The algorithm comprises a set of empirical rules dividing chemical compounds into retrosynthetically meaningful substructures. A substructure can belong to one of 16 predefined classes. Fig. 1 shows an example where each color on an atom represents the assignment to a specific BRICS class. Substructures span a variable number of atoms and their assignment is highly dependent on the overall chemical structure. We therefore postulate that a node-level classification task based on the BRICS assignment is a valuable pre-training task. To the best of our knowledge this is the first application of the BRICS algorithm as a pre-training task for graph neural networks with chemical structure as inputs. For training on the above tasks we use the ChEMBL database, leveraging roughly 2.1mio bio-active compounds.

In addition to the self-supervised approaches we also experiment with training using labeled in-house data where we expect a positive correlation to our downstream human PK prediction tasks. To that end, two datasets are used. The first set is constructed from human in vitro experiments containing roughly 200k unique structures. Training labels include experiments for solubility in 50 mM phosphate buffer at pH 6.5 (n=110k), lipophilicity (n=102k), membrane permeability (n=109k), P-glycoprotein (Pgp) apical efflux ratio (n=6k), protein binding affinity expressed as fraction unbound in human plasma (n=14k), as well as microsomal (n=94k) and hepatic (n=11k) clearance. The second set comprises mouse and rat PK parameters and contains roughly ten thousand unique structures. Details regarding the rodent PK dataset have previously been published by Stoyanova et al. For both labeled datasets we test two training modes: A pre-training mode where the network is first trained on the upstream dataset and then fine-tuned on the downstream set in a separate optimization procedure (Fig. 2d,e). Alternatively, up- and downstream datasets can be combined and the model is trained to predict all labels simultaneously (Fig. 2b,c). For all pre-trained models (Fig 2, second row), the message-passing part of the network is trained at a reduced learning rate of $5 \times 10^{-5}$ when fine-tuning the model to the downstream human PK prediction task. The network’s remaining feed forward parts are trained at the initial learning rate of $1 \times 10^{-3}$. Figure 2 summarizes all training schedules and the respective data sets used.

**Model performance metrics**

Following our earlier work on the prediction of rodent PK properties we evaluate the predictive quality of our models considering accuracy, bias and correlation. To do so we introduce a minimal amount of mathematical notation: All equations below assume that $\mathbf{y} = (y_1, ..., y_n)$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, ..., \tilde{y}_n)$ are vectors of reference and predicted values, respectively, each holding $n$ elements. Note that our models predict endpoints in the logarithmic space (see Dataset cu-
Figure 2: Overview of all training schedules and the respective datasets used in this work. Top row shows schedules where a multi-task model is trained once to predict the labels of one or multiple combined data sets. Bottom row shows approaches that employ pre-training followed by fine-tuning. For all data sets a rounded number of unique chemical structures is printed in the second row. Green and pink colors indicate primary and auxiliary resources, respectively.

ration & analysis Section) whereas some of the metrics operate on the real space. For a log_{10}-transformation this implies that the real space values are recovered through \( y = 10^{y_{log}} \), where \( y_{log} \) are the values in logarithmic space. When evaluating the accuracy, a standard metric in the scientific literature related to PK modeling is the geometric mean fold error (\( GMFE \)), otherwise known as the absolute average fold error (\( AAFE \))

\[
GMFE(y, \hat{y}) = \\
AAFE(y, \hat{y}) = \exp\left(\frac{1}{n} \sum_{i=1}^{n} |\ln \frac{\hat{y}_i}{y_i}|\right) \tag{4}
\]

The \( AAFE \in [1, \infty) \) can be interpreted as the average factor by which the predictions are off from the true values, without distinction between over- or underprediction, with an ideal value of 1. To measure prediction bias the average fold error (\( AFE \)) is used:

\[
AFE(y, \hat{y}) = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \ln \frac{\hat{y}_i}{y_i}\right) \tag{5}
\]

The \( AFE \in (0, \infty) \) has an ideal value of 1, with smaller and larger values indicating under- and overprediction, respectively. Although domain experts might be most familiar with the above metrics, we would like to advise caution when using them to compare different models that predict properties in the logarithmic space: Because back-transformation to the real space (e.g. \( y = 10^{y_{log}} \)) is non-linear, the metrics cannot be compared reliably across different data sets or PK endpoints. For example, assuming a model with a constant level of accuracy for all predicted endpoints the \( AAFE \) will be naturally larger for properties where the range of experimental data spans multiple orders of magnitude. Therefore, when comparing predictive performance of various models, different (\( AFE \) values can be compared reliably only if the reference data \( y \) comes from the same – or otherwise very similar – distribution. We argue that comparing metrics in the training data space is more meaningful as it avoids the above effect. For this reason, we include the logarithmic root-mean-square error

\[
RMSE_{log}(y_{log}, \hat{y}_{log}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{log,i} - \hat{y}_{log,i})^2} \tag{6}
\]

as an additional metric for model accuracy, with an ideal value of 0 and a range of
$RMSE_{\log} \in [0, \infty)$. Similarly, linear correlation between the ground truth and predicted values is measured with the Pearson Correlation Coefficient

$$\rho_{\log}(y_{\log}, \hat{y}_{\log}) = \frac{\text{cov}(y_{\log}, \hat{y}_{\log})}{\sigma_{y_{\log}} \sigma_{\hat{y}_{\log}}} \quad (7)$$

where $\text{cov}(\cdot, \cdot)$ and $\sigma$ denote covariance and standard deviation, respectively. Values for $\rho$ are in the range of $\rho \in [-1, 1]$ where 1 indicates perfect linear correlation.

Epistemic uncertainty quantification

During the early stages of drug development, design teams tend to explore large and novel chemical spaces. This stands in contrast to a comparably small and slowly-expanding data set of clinical compounds, where patents further limit the presence of chemically similar structures. We therefore expect the accuracy of our human PK models to be highly variable, as tested chemical compounds will vary in their similarity to the training data. To mitigate this, we consider a model’s ability to estimate meaningful epistemic uncertainties (i.e. the model uncertainty stemming from the lack of data) a crucial functionality. We apply and compare the following methods for the estimation of model uncertainty in this work:

- Deep neural network ensembles.\textsuperscript{46} Multiple instances of the same architecture are trained. Prior to training the weights of each model instance are initialized to random numbers, representative of an arbitrary starting position in the parameter search space. The technique assumes that by training several independent, randomly initialized models on the same data, the variability of predictions in regions of sufficient training data will be lower compared to the data-scarce regions.
- Evidential deep learning.\textsuperscript{47–49} In the case of regression tasks, the approach assumes a normal-inverse-gamma prior for each point in the training data. An evidential model learns to predict the parameters of the aforementioned distribution with the promise to model both epistemic (model) and aleatoric (data) uncertainty in addition to the mean.
- Mean-variance estimation (MVE).\textsuperscript{50,51} Similar to evidential deep learning, it assumes a Gaussian prior on each data point, thus effectively modeling the mean and variance of the distribution.
- Quantile regression.\textsuperscript{52,53} Rather than estimating the mean of a conditional distribution, quantile regression is a method for estimating any arbitrary quantile of the training data. In practice this is achieved by utilizing the Pinball Loss (otherwise known as the Quantile Loss) function during training. The loss can be interpreted as an extension of the mean absolute error, adapted to minimize any arbitrary quantile. Here we fit our models to predict the quantiles associated with the 5% and 95% probabilities.
- Gaussian process regression (GPR).\textsuperscript{51,54} Assumes a multivariate Gaussian prior and fits a parametrized covariance function, conditioned on the training data to model the posterior distribution. We use the vectors from the penultimate layer of our neural network model as the input features to the GPR. To that end we use the GPyTorch package\textsuperscript{55} (v1.11) with a radial basis function kernel as the covariance function. Note that the GPR-estimated means are discarded and only the variances are used.

All uncertainty quantification (UQ) methods were applied on the GPS model.

Evaluating epistemic uncertainty

Our primary requirement for the epistemic uncertainties produced by a model is that they are accurate with respect to the detection of out-of-distribution inputs. We therefore expect a correlation between prediction error and the magnitude of the returned uncertainty. To the model user this translates to the ability
to reliably tell to what degree model predictions can be trusted and act accordingly on this additional information. To the best of our knowledge there appears to be no established standard to evaluate this property, which we will refer to as ‘confidence’, in the scientific literature. Common to the field of optical flow measurement are sparsification plots, otherwise known as confidence curves.\textsuperscript{56,57} They are a visual method used for the evaluation of a model’s confidence. An application in the context of predictive models in chemistry has recently been proposed by Scalia and coworkers.\textsuperscript{58} Confidence curves are generated as follows. Given a set of test data points with reference values, predicted values, and prediction uncertainties, a fraction $p$ of all entries is removed from the complete set in the order of increasing uncertainties. For the remaining set an average error metric is calculated based on the true and predicted values. A confidence curve is generated by selecting a number of fractions $p \in [0, 1]$, and plotting the respective error at that fraction. At $p = 0$ the complete set is considered, whereas at $p = 1$ all points are removed and the error is zero. An example for three confidence curves is depicted by the green lines in Figure 3. We can repeat the above process, but instead of sorting by increasing uncertainty we sort by the error itself. In doing so we retrieve a curve associated with the ideal sorting, which is depicted by the red line in Figure 3. Once both curves are obtained, the area between the two curves (Fig. 3 grey area) can be used as a numerical measure for the confidence of a model.

Using the area as defined above is problematic because (1) the ideal curve will always vary with the data set and model performance thus making a direct comparison impossible, and (2) there are no independent anchors to the numeric value of the area, which makes a qualitative discussion difficult. Here we propose a solution to both issues. First, we normalize all partial error metrics by the complete set’s error, effectively making the error at $p = 0$ unity. Second, we normalize the area between the two curves to make it independent of the ideal curve’s shape. If $A$ and $B$ are the integrals of the ideal and the uncertainty-ordered curves, Figure 3: Three examples of a confidence curve for the case of a positive correlation (a), no correlation (b), and a negative correlation (c) between magnitude of uncertainties and model error, respectively. The green line indicates how a confidence curve could look like if entries were removed in the order given by the magnitude of the individual uncertainty values. The red dotted line indicates the ideal confidence curve – a lower bound resulting from the removal order given by the actual error metric itself. Grey area between the curves marks the area under the confidence curve ($\text{AUCC}$).
respectively, a normalized area under the confidence curve (AUCC) can be computed as

\[ \text{AUCC} = \frac{B - A}{1 - A} \] (8)

Intuitively, AUCC = 1 when the model uncertainty is on average not correlated with the prediction error (Fig. 3b). A model is confident when AUCC < 1, where smaller values mean increased confidence (Fig. 3a). Conversely, an AUCC > 1 indicates a negative correlation between uncertainty and error (Fig. 3c).

The AUCC is a ranking-based metric that does not consider the magnitude of the returned uncertainties. We therefore introduce calibration as a secondary property when evaluating model uncertainty. A common metric for model calibration is the expected calibration error (ECE).\(^{59,60}\) This assumes that a model’s prediction and uncertainty at a given point can be interpreted as parameters describing a Gaussian distribution. Under this assumption, given a confidence level \(q \in [0, 1]\), we can construct a confidence interval for each point in a test data set. A model is well-calibrated if for all \(q\), the fraction of points for which the confidence interval covers the true labels, \(q_{\text{obs}}\), is equal to \(q\). For a given set of confidence levels \(Q = \{q_0, \ldots, q_n\}\) the ECE can be computed as the mean absolute deviation between all \(q_i\) and \(q_{i, \text{obs}}\):

\[ \text{ECE} = \frac{1}{n} \sum_{i=1}^{n} |q_{i, \text{obs}} - q_i| \] (9)

The metric has a range of \(\text{ECE} \in [0, 1]\), where lower values indicate a better calibration. Here we choose \(Q\) to contain 30 equidistant confidence levels in the complete range. Because the ECE is calculated as a mean across the whole data set, a model can be well-calibrated on average without any guarantees to smaller subsets of data. Levi and coworkers therefore measure calibration with the expected normalized calibration error ENCE.\(^{60}\) The ENCE interprets uncertainty as the variance of a Gaussian distribution and assumes that for each point, the ratio between the root-squared error and the squared variance is close to one in a well-calibrated model. To calculate the ENCE in practice, we sort our test set residuals by increasing uncertainty (as with the AUCC) and divide it into \(n = 30\) bins. For each bin, we calculate the \(\text{RMSE}\) and the root mean variance \(\text{RMV} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \sigma_i^2}\) (\(m\) is the number of entries per bin). The \(\text{ENCE} \in [0, \infty)\) is then calculated as

\[ \text{ENCE} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\text{RMV}_i - \text{RMSE}_i|}{\text{RMV}_i} \] (10)

While not in scope of this work, note that all models can be re-calibrated towards lower \(E(N)CE\) values.\(^{59,60}\) We therefore consider confidence and the related AUCC a more relevant metric when evaluating uncertainty. All uncertainty metrics are evaluated in the logarithmic PK endpoint space.

Results

Human pharmacokinetics dataset

To the best of our knowledge this work curates and utilizes the largest data set of clinical pharmacokinetics described in the scientific literature to date. Shedding light on the current landscape of clinical PK for small molecules (molar weight \(\leq 1000\) g/mol), we present a comprehensive analysis of the data set below. We summarize the curated data in Table 1, listing relevant statistical descriptors of all parameter value distributions. Additional data allowing for the exhaustive comparison of CDDI and PP data is provided in the Supporting Information (Sec. S2) in the form of tables, histogram plots for all parameters, as well as heatmaps depicting the overlap of entries for any two PK parameters. The data set is highly diverse, covering a broad experimental range for all pharmacokinetic endpoints. The same is true for common physico-chemical properties, expressed through estimates of the octanol-water partition coefficient (logP), polar surface area and the molecular weight. While distribution of logP is close to Gaussian, covering over two orders of magnitude, long tails in the higher val-
ues are observed for the latter two descriptors. We find that both data sets are similar and well within Lipinski’s rule of five, with CDDI containing slightly smaller and more hydrophilic compounds on average. The mean number of hydrogen bond acceptors and donors is 5.5 and 2, respectively. Considering the individual PK endpoints, we observe significant overlap between CDDI and PP. For both sources a higher number of PK parameters associated with the oral route of administration is present. In the case of CDDI data no distinction between clearance (CL), volume (V), and the apparent counterparts – CL/F, V/F is made. Due to this ambiguity we excluded all CDDI data for the apparent parameters. For parameters where both sources are considered CDDI consistently provides more data. This is in line with the broader scope of sources used by CDDI, which in contrast to PP also include studies that are not part of official approval packages for government authorities. To establish a lower bound for the accuracy of each PK endpoint, estimates of experimental variability – expressed through the absolute average fold error (AAFE) of multiple measurements in relation to their median – are included in Table 1. Overall, the results align with the common findings in experimental pharmacokinetics: Depending on the parameter, the AAFE ranges between 1.1 and 1.5 with errors for oral administration generally higher than their intravenous counterparts.

Turning to the chemical space analysis, we report a unique chemical structure count of 905 and 2495 for PP and CDDI, respectively. Set similarity coefficients (see Chemical space analysis Section) and the exact overlap of chemical structures are reported in Table 2. In addition to CDDI and PP, we also compare both sets to (1) a data set previously published by Lombardo et al., which includes some 1300 drug compounds from the clinic, and (2) an in-house data set of compounds that were part of a PK study in rodents. The latter contains roughly ten thousand unique structures and is described in more detail by Stoyanova et al. Similarity coefficients between the three data sets of clinical compounds are relatively high, which is in part also due to the high overlap of exact chemical compound matches. Conversely, comparison with the rodent PK data set shows a significantly smaller similarity coefficient of roughly 0.3 which highlights the considerably different nature of compounds in drug discovery compared to the clinic.

Although the similarity coefficient is helpful as a global measure, it does not capture local variations of the compared distributions. We therefore consider a visual depiction of the chemical space covered by the PP and CDDI data sets in the form of a UMAP embedding in Figure 4. The top left part of the figure shows how structures from CDDI (blue points) and PP (red points) are distributed across a reference space, which is depicted by the gray point cloud. The latter is constructed from a random selection of 250 thousand structures from the ChEMBL database (see Chemical space analysis Section). On a local scale all three datasets exhibit clusters of higher densities at similar positions in the embedded space, supporting the results of high set similarity coefficients. Examples of structures from five higher-density regions are provided in the lower left part of the figure. We further note that both CDDI and PP have a good overall coverage of the complete ChEMBL reference space. Such congruence can be explained the ChEMBL library mostly capturing drug-like and bio-active molecules.

Availability of indications per compound in the CDDI database allows us to map most structures to a set of corresponding ICD-10 (10th revision of the International Statistical Classification of Diseases and Related Health Problems) chapters. The most prevalent chapters, based on the relative percentage of compounds with indications that match this chapter are

- I: Infectious and parasitic diseases (25.7%)
- II: Neoplasms (24.8%)
- IX: Diseases of the circulatory system (24.0%)
- VI: Diseases of the nervous system (22.8%)
- V: Mental and behavioural disorders (22.6%)
Table 1: Statistical descriptors of the human PK data set. Showing the number of entries (structure and dose), number of unique SMILES (i.e. structures), lower and upper quartiles (Q1, Q2) as well as mean and median values and the experimental estimate of the absolute average fold error ($AAFE_{exp}$), per PK property. Dose statistics are calculated for the complete set. Additionally showing descriptors for the molar weight (Molwt), polar surface area (PSA) and the octanol-water partition coefficient (logP) for all unique compounds.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$N_{Entries}$</th>
<th>$N_{SMILES}$</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>$AAFE_{exp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC$_{IV}$ [mol h/L]</td>
<td>2118</td>
<td>683</td>
<td>4.58e-7</td>
<td>6.40e-6</td>
<td>3.55e-4</td>
<td>7.15e-5</td>
<td>1.36</td>
</tr>
<tr>
<td>AUC$_{PO}$ [mol h/L]</td>
<td>8259</td>
<td>1969</td>
<td>6.67e-7</td>
<td>4.91e-6</td>
<td>1.38e-4</td>
<td>3.25e-5</td>
<td>1.48</td>
</tr>
<tr>
<td>CL [L/h]</td>
<td>2170</td>
<td>727</td>
<td>7.37</td>
<td>1.91e1</td>
<td>5.24e1</td>
<td>4.54e1</td>
<td>1.27</td>
</tr>
<tr>
<td>CL/F [L/h]</td>
<td>991</td>
<td>318</td>
<td>6.79</td>
<td>1.97e1</td>
<td>1.21e2</td>
<td>6.50e1</td>
<td>1.33</td>
</tr>
<tr>
<td>C$_{max}$ [mol/L]</td>
<td>9783</td>
<td>2211</td>
<td>7.26e-8</td>
<td>5.93e-7</td>
<td>9.27e-6</td>
<td>3.35e-6</td>
<td>1.51</td>
</tr>
<tr>
<td>F [%/100]</td>
<td>1167</td>
<td>577</td>
<td>2.48e-1</td>
<td>4.88e-1</td>
<td>4.90e-1</td>
<td>7.50e-1</td>
<td>1.13</td>
</tr>
<tr>
<td>T$_{half,IV}$ [h]</td>
<td>2255</td>
<td>783</td>
<td>1.77</td>
<td>4.00</td>
<td>1.30e1</td>
<td>1.10e1</td>
<td>1.24</td>
</tr>
<tr>
<td>T$_{half,PO}$ [h]</td>
<td>8212</td>
<td>2067</td>
<td>4.11</td>
<td>9.30</td>
<td>1.91e1</td>
<td>1.97e1</td>
<td>1.30</td>
</tr>
<tr>
<td>T$_{max}$ [h]</td>
<td>9402</td>
<td>2184</td>
<td>1.10</td>
<td>2.00</td>
<td>2.85</td>
<td>3.55</td>
<td>1.37</td>
</tr>
<tr>
<td>V [L]</td>
<td>1879</td>
<td>682</td>
<td>2.12e1</td>
<td>7.45e1</td>
<td>2.52e2</td>
<td>2.18e2</td>
<td>1.23</td>
</tr>
<tr>
<td>V/F [L]</td>
<td>698</td>
<td>243</td>
<td>8.30e1</td>
<td>2.78e2</td>
<td>1.51e3</td>
<td>1.11e3</td>
<td>1.23</td>
</tr>
<tr>
<td>Dose [g]</td>
<td>46934</td>
<td>2702</td>
<td>1.00e-2</td>
<td>5.86e-2</td>
<td>2.23e-1</td>
<td>2.50e-1</td>
<td></td>
</tr>
<tr>
<td>Molwt [g/mol]</td>
<td>2702</td>
<td>301</td>
<td>388</td>
<td>400</td>
<td>474</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSA [Å$^2$]</td>
<td>2702</td>
<td>5.86e1</td>
<td>8.36e1</td>
<td>9.00e1</td>
<td>1.11e2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>logP</td>
<td>2702</td>
<td>1.64</td>
<td>2.97</td>
<td>2.88</td>
<td>4.24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Set similarity coefficients $S$ for the PP, CDDI, Lombardo$^3$ and rodent$^4$ PK data sets pairs. Number of same compounds in both sets in parenthesis.

<table>
<thead>
<tr>
<th></th>
<th>CDDI</th>
<th>Lombardo</th>
<th>Rodent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>0.61 (806)</td>
<td>0.60 (473)</td>
<td>0.31 (75)</td>
</tr>
<tr>
<td>CDDI</td>
<td>0.57 (732)</td>
<td>0.33 (138)</td>
<td>0.29 (64)</td>
</tr>
</tbody>
</table>

- IV: Endocrine, nutritional and metabolic diseases (18.1%)
- XV: Pregnancy, childbirth and the puerperium (0.4%)
- XVII: Congenital malformations, deformations and chromosomal abnormalities (0.6%)
- VIII: Diseases of the ear and mastoid process (0.9%)
- VII: Diseases of the eye and adnexa (7.4%)
- XII: Diseases of the skin and subcutaneous tissue (8.8%)

Highlighting all compounds that match the respective chapter in the UMAP embedding space reveals some interesting patterns, presented in the smaller subfigures in the lower part of Figure 4. For example, the island on the left in the embedding space is mostly occupied by compounds belonging to chapters I and II. Compounds associated with chapters IV and V are preferentially located in the right part of the embedding. On average, a compound’s indication covers 2.4 ICD-10 chapters. Chapters with the least prevalence are

- XV: Pregnancy, childbirth and the puerperium (0.4%)
- XVII: Congenital malformations, deformations and chromosomal abnormalities (0.6%)
- VIII: Diseases of the ear and mastoid process (0.9%)
- VII: Diseases of the eye and adnexa (7.4%)
- XII: Diseases of the skin and subcutaneous tissue (8.8%)

The spatial UMAP embedding is combined with key physicochemical properties in Section S1.5 of the SI, showing all embedded structures as points, colored by molar weight, lipophilicity and polar surface area. This allows us to observe that structures with a lower polar surface area are mostly located in the lower-right of the UMAP embedding space. Compounds with a higher lipophilicity can be found in the left and right of the embedding, but not in the middle part. Heavier compounds tend to be located at the borders of the clusters throughout the complete space except the top-right part of the embeddings, where the heavier compounds...
are predominant.

**Model development**

The following presents a summary of all cross-validation experiments leading up to the design of our best-performing model. In this section, whenever we speak of significant results, we refer to the outcome of a two-sided t-test with a p-value below 0.05. For clarity, we restrict ourselves to a qualitative description in the main text. A more comprehensive comparison of all experiments in the form of performance metric tables per PK endpoint – with significant results denoted – is provided in the Supporting Information (Sec. S2).

**Prediction accuracy and correlation**

We provide the average cross-validation performance per endpoint, expressed through the absolute average fold error (AAFE), average fold error (AFE), logspace root-mean square error (RMSE\textsubscript{log}) and logspace Pearson’s correlation coefficient (\(\rho_{\text{log}}\)) in Section S2 of the SI. All experiments are divided into three groups (as denoted by the leading number in each table).

Prior to experimenting with different models, we aimed to establish how different data splitting techniques affect model performance in Group (1). To that end a static (i.e. same configuration and training approach) GPS model was used. Overall the cross-validation strategy does not significantly affect model performance. Although random splitting on average results in slightly better AAFE and \(\rho_{\text{log}}\) values, the effect is less pronounced when turning to the AFE and RMSE\textsubscript{log} metrics. The lack of significance suggests that the data set we are dealing with is highly diverse, as random split performance is comparable to the SCINS\textsuperscript{26,64} split. We see this explained through our previous findings, which indicate that the overall chemical space of the human PK data is highly diverse and mostly sparse. Due to the insignificant differences in model performance, we use the random splitting strategy for all following cross-validation experiments.

Experiment Group (2) compares the predictive performance across different model architectures. We observe that training graph neural network-based architectures like our GPS model in a multi-task fashion is beneficial to the overall model performance. Comparing across the GPS, Chemprop and XGBoost model architectures the overall performance is similar, with correlations ranging from strong (\(\rho_{\text{log}} > 0.6\), V, \(C_{\text{max}}\), AUC\textsubscript{PO}, AUC\textsubscript{IV}) to mediocre (0.6 > \(\rho_{\text{log}} > 0.3\), all remaining endpoints). AAFE metrics vary by endpoint, with a value below 3 for CL, F, \(T_{\text{max}}\), \(T_{\text{half,PO}}\), \(T_{\text{half,IV}}\) and V. The remaining endpoints are predicted with an AAFE ranging from 3.2 (V/F) to 4.3 (AUC\textsubscript{PO}). Within the group, XGBoost performs significantly better than (AUC\textsubscript{PO}, \(C_{\text{max}}\)) or on par with (remaining endpoints) the GNN-based architectures, suggesting that GNN performance could be further improved, either through more elaborate engineering of input features or a more exhaustive search for a better model configuration. Given the scope of this work we do not pursue the search for a marginally more accurate baseline model, and instead focus on the exploration of general trends given a static model configuration. A clear advantage of neural networks over more classical models like XGBoost is their ability to handle sparse multi-labelled data and the option of parameter fine-tuning. As a lower boundary of model performance we additionally include trivial predictors where each value in the test set is predicted as the mean or median value in the training set.

Considering the AAFE, we observe that especially when using the median of a distribution as an estimator, there is no significant difference to the trained machine learning models for the following parameters (median-estimated AAFE in parentheses): CL/F (4.2), CL (3.0), F (2.2), \(T_{\text{max}}\) (1.8), \(T_{\text{half,IV}}\) (3.2) and V/F (3.8). Generally this effect can be observed wherever the experimental range of a PK property is small (up to one order of magnitude), with median estimators for AUC\textsubscript{IV}, AUC\textsubscript{PO} and \(C_{\text{max}}\) performing significantly worse (AAFE > 9.0). Assuming that train and test set values are independent and identically distributed, using the mean or median will result in an unbiased esti-
Figure 4: Top: UMAP embedding of the CDDI (blue) and PP (red) chemical structures into a two-dimensional space fitted to a random selection of 250 thousand compounds ChEMBL library (grey). Insets in the lower part of the UMAP depict individual embeddings per data source for reference. Four representative structures from the human PK data set are drawn at different higher-density locations of the embedding space, as depicted by the cyan rectangles labeled A-E and depicted on the right part of the figure. Bottom: UMAP embeddings where all compounds in the human PK data matching one of the top 8 most-prevalent ICD-10 chapters – (a)-(h), respectively – are highlighted in blue.
mation ($AFE \approx 1$). As expected there is very little correlation between predicted and experimental values for naive constant-value estimators. This discrepancy between a good $AAFE$ and poor correlation metrics for trivial estimators highlights the need for the evaluation of multiple metrics to fully capture model performance.

GPS performance after application of training strategies that involve additional data is presented in Group (3). Considering the self-supervised pre-training methods MolCLR\textsuperscript{44} and BRICS, we observe a clear loss in performance compared to the baseline. Moreover, similar to the trivial estimators, endpoints where the experimental data spans a larger range perform worse than the ones with a smaller range. One explanation for this outcome might be that the pre-training objectives are too simple and as a result do not transfer to the complex task of human PK predictions. On the other hand, leveraging additional labeled data sets appears to be beneficial for the performance of the downstream prediction tasks. Following the experimental drug design process, incorporation of both rodent PK and \textit{in vitro} data sets into our training pipeline improves the predictions of human PK. For training with these two auxiliary data sets, we distinguish between two modes: A pre-training/fine-tuning approach where the model is first trained on the auxiliary data set and the same parameters are used as a starting point for training on the human PK afterwards; and a multi-task approach, where the auxiliary data set is combined with the human PK data set and the model is trained only once to predict all endpoints. Models generated with the latter strategy are denoted as ‘MT’ in the SI tables and appear to perform slightly better than their pre-trained counterparts. We observe a larger improvement in the human PK predictions when training in combination with \textit{in vitro} data as one data set in a multi-task fashion. There is no significant difference in the chemical structure-based splitting approaches, which is likely owed to the small size and large diversity of the data set. Finally, we identify a number of difficult endpoints, namely clearance, bioavailability as well as elimination half-time and time and maximum concentration. Due to the small experimental range of these endpoints low prediction accuracy values can be achieved easily and model evaluation warrants a cautious comparison to meaningful baselines.

**Epistemic uncertainty quantification**

We first turn to the discussion of the single-method approaches for uncertainty quantification as presented in the Epistemic uncertainty quantification Section, namely, mean-variance estimation (MVE), Gaussian process regression (GPR), quantile regression (QR), evidential deep learning and ensembling (please refer to the SI Sec. S2.5-S2.7 for tables showing the averaged AUCC, ECE and ENCE metrics). Considering these five, MVE performs best across most PK endpoints when measured with the AUCC. Selection of a second-best method is not straightforward as none appears to be significantly better among the remaining ones. With regards to the calibration metrics ECE and ENCE both MVE and GPR are well-calibrated, whereas calibration varies significantly per endpoint for the evidential deep

https://doi.org/10.26434/chemrxiv-2024-vgbxq ORCID: https://orcid.org/0000-0001-6011-1632 Content not peer-reviewed by ChemRxiv. License: CC BY-NC-ND 4.0
learning method. Uncertainties from ensembling and QR methods are not well-calibrated. While in the case of the former this is a known issue, poor calibration of QR is due to the fact the method does not assume a Gaussian prior on our model predictions and therefore invalidates the assumptions of both calibration metrics.

Performance of individual PK endpoints is somewhat variable. Model confidence is very low for bioavailability (F) and $T_{\text{max}}$ with an AUCC close to 1 and rather poor calibrations ($ECE > 0.2$). While other endpoints perform better, the lowest observed value of AUCC = 0.67 suggests that establishing meaningful uncertainties in the context of clinical PK predictions is an overall challenging task.

Following up on approaches that leverage only one method for the quantification of uncertainty, we show that combining ensembling with other approaches can improve the overall quality of the returned uncertainties. Lower AUCC values are observed when applying ensembling to MVE, effectively evaluating the average mean and variance of eight independently-trained models (denoted ‘MVE ens.’ in the SI tables). Building on the findings from the previous section we train an MVE ensemble on a multi-task data set that contains in vitro data in addition to the human PK endpoints (denoted ‘MVE ens. ivMT’ in the SI tables). Combination of MVE, ensembling and the training on auxiliary in vitro tasks produces the most meaningful uncertainties with respect to the AUCC across all human PK endpoints while retaining a good level of calibration.

To highlight how meaningful uncertainties translate to a practical application we once again consider model accuracy and correlation Section S2.8 of the SI, comparing the baseline GPS model to our best-performing model GPS*, which is an ensemble of MVE models trained on in vitro and human PK data in a multi-task fashion. GPS* outperforms the baseline across all tasks and metrics. More interestingly, access to prediction uncertainties allows us to reliably identify subsets of data that are more or less accurate than the complete set. Based on the uncertainty rankings, we select the most and least confident 20% of all predictions and compare their average performance in the same table below. The results show that this approach allows us to identify regions of better or worse predictive performance. Moreover, we see that the approach is more reliable for endpoints with a lower AUCC, underlining the validity of the metric. We perform and discuss the same analysis on the test set predictions in the following section.

**Test set performance**

After selecting the best model based on the preceding cross-validation experiments, we report the predictive performance on previously unseen data in this section. The test set was selected by setting aside ten percent of the initial human PK data, based on a randomized choice of unique chemical structures. While a random split can produce overly-optimistic results in a setting where a larger number of chemically similar compounds is present (cf. e.g. Stoyanova et al.4), our cross-validation experiments indicate that this effect is negligible in the case of a small and diverse chemical space (see Model development Section). The test set contains a total of 4481 entries and 270 unique chemical structures. Individual numbers per PK endpoint are depicted in Figure 5. In S3.1, we compare the parameter distributions between training and test set as box plots for each PK parameter. Overall little difference in the distributions is observed when comparing the two sets.

We summarize all model performance metrics in Table 3, showing the $AAFE$, $RMSE_{\log}$, fraction within 2-fold error, Pearson correlation coefficient $\rho_{\log}$ and $AFE$ for each PK endpoint. In addition, Figure 5 displays all individual ground truth values against their predictions. For the uncertainty metrics as well as plots of the underlying confidence curves per endpoint please refer to the Supporting Information Section S3. Performance trends are mostly in line with the results from cross-validation experiments. Comparing model accuracy, we observe some variability between the individual endpoints, with the $AAFE$ ranging from 4.5
(AUCPO) to 1.7 (Tmax). We are pleased to observe that seven out of eleven predicted properties can be predicted with an AAFE < 3, namely clearance (CL), bioavailability (F), time at maximum concentration (Tmax), elimination half times (Thalf,IV, Thalf,PO) and volumes (V, V/F). Strong correlations of plog > 0.7 can be observed for AUCIV, AUCPO, Cmax and V/F. Correlations for the remaining parameters are in the range of 0.48 (Tmax) and 0.6 (Thalf,IV, Thalf,PO, CL/F). Turning to the AFE, our model appears to be biased towards under-predicting AUCIV (0.69), AUCPO (0.57) and Cmax (0.69). Over-prediction bias is observed for CL/F (1.42), F (1.39) and V/F (1.33). The remaining properties – CL, Tmax, Thalf,IV, Thalf,PO and V – do not exhibit a strong bias. Results in the table additionally illustrate how uncertainty quantification can be applied to detect a model’s applicability domain. We use predicted model uncertainties to derive two subsets from the complete test data: Twenty percent of all entries where our model is the most and least confident, denoted as ‘Top 20%’ and ‘Bottom 20%’, respectively. Recalling our requirement that small uncertainties are correlated with small errors we expect that for the Top 20% subset of the test data, metrics are significantly improved over the complete set. We see this requirement largely met when comparing the subset’s metrics: All but one endpoint (AUCIV) are predicted with an AAFE < 3, with eight out of eleven endpoints within an AAFE < 2.5. Likewise an improvement in correlation is observed for most endpoints, with the exception of F and AUCIV. Poor model confidence for bioavailability was expected based on our previous cross-validation results (AUCC ≈ 0.9, see S2.5). In contrast to accuracy and correlation, no clear claim of improved performance with respect to model bias can be made. When considering the AFE, we do not observe a distinct separation between the complete data and the respective subsets. The ability to confidently detect better subsets does not guarantee the inverse, namely the ability to detect subsets that are worse with the same level of confidence. For this reason metrics for the bottom 20% of the complete test set are likewise included in Table 3. We indeed observe worse-than-complete-set performance for all but two endpoints, indicating that detection of poorly- and well-described regions works equally well. Exception to this are the volume endpoints, where metrics for the bottom 20% are better than for the complete test set.

Discussion

Expressions such as ‘predictions of clinical pharmacokinetics from chemical structure alone are highly challenging’ have become somewhat of a mantra in the scientific literature when it comes to managing expectations. Generally, our findings support this perspective. At the same time, given that the field of pharmacokinetic modeling has only recently started to warm up to data-driven approaches such as machine learning, we are confident that there is plenty of room to push the boundaries. In that spirit, we hope that this work can contribute its part by highlighting novel developments and possibilities.

Collection and curation of data for the training of our models has been a major contribution of this study. To the best of our knowledge our study comprises the largest and most diverse data set of clinical PK described in the scientific literature to date, covering over 2700 compounds across 11 PK parameters for the intravenous and oral routes of administration. While we can not release the data set due to contractual constraints, we are confident that the exhaustive analysis and reported statistics will be of significant interest to the community, providing a recent image of the clinical PK space for small molecules. The extracted data can not be considered free of experimental noise due to the use of different protocols, equipment, human error, and possible errors in the automated extraction of the data from literature. We aimed to provide an estimate of this error in the calculation of an experimental AAFE in Table 1.

To analyze and compare the chemical similarity of multiple sets two approaches were employed. For a quantitative evaluation of two sets of chemical compounds we proposed adapt-
Figure 5: Test set correlation plots showing experimental vs. predicted values for each endpoint. Subset identified as the top 20% of the most confident predictions is highlighted in green. Dashed diagonal lines indicative of $AAFE = 2, 3$. Numbers below each parameter are the total number of entries and unique chemical structures, respectively.
Table 3: Test set accuracy and correlation metrics per endpoint, as predicted with GPS*. In addition to evaluating the complete test set, metrics for two subsets are shown. Selection is based on the most confident (top 20%) and least confident (bottom 20%) predictions according to the model uncertainty. Bold and underlined values mark the best and second-best results, respectively. Grey values indicate performance worse than the complete test set.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Target Test set</th>
<th>AUC IV</th>
<th>AUC PO</th>
<th>CL/F</th>
<th>CL</th>
<th>C_max</th>
<th>F</th>
<th>T_max</th>
<th>T_half,IV</th>
<th>T_half,PO</th>
<th>V/F</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAFE</td>
<td>Complete</td>
<td>4.48</td>
<td>3.85</td>
<td>3.28</td>
<td>2.49</td>
<td>3.33</td>
<td>2.15</td>
<td>1.65</td>
<td>2.36</td>
<td>2.05</td>
<td>2.92</td>
<td>2.45</td>
</tr>
<tr>
<td></td>
<td>Top 20%</td>
<td>3.94</td>
<td>2.84</td>
<td>2.87</td>
<td>2.18</td>
<td>2.23</td>
<td>2.37</td>
<td>1.53</td>
<td>1.42</td>
<td>1.77</td>
<td>1.69</td>
<td>1.84</td>
</tr>
<tr>
<td></td>
<td>Bottom 20%</td>
<td>6.62</td>
<td>4.87</td>
<td>5.61</td>
<td>2.67</td>
<td>5.69</td>
<td>3.06</td>
<td>1.83</td>
<td>3.05</td>
<td>2.56</td>
<td>2.17</td>
<td>2.22</td>
</tr>
<tr>
<td>RMSE_{log}</td>
<td>Complete</td>
<td>0.96</td>
<td>0.75</td>
<td>0.67</td>
<td>0.63</td>
<td>0.72</td>
<td>1.65</td>
<td>0.28</td>
<td>0.52</td>
<td>0.41</td>
<td>0.6</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Top 20%</td>
<td>0.87</td>
<td>0.58</td>
<td>0.54</td>
<td>0.49</td>
<td>0.49</td>
<td>1.82</td>
<td>0.24</td>
<td>0.18</td>
<td>0.32</td>
<td>0.33</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Bottom 20%</td>
<td>1.22</td>
<td>0.88</td>
<td>0.92</td>
<td>0.58</td>
<td>1.02</td>
<td>1.83</td>
<td>0.32</td>
<td>0.58</td>
<td>0.5</td>
<td>0.47</td>
<td>0.45</td>
</tr>
<tr>
<td>ρ_{log}</td>
<td>Complete</td>
<td>0.74</td>
<td>0.8</td>
<td>0.6</td>
<td>0.54</td>
<td>0.81</td>
<td>0.54</td>
<td>0.48</td>
<td>0.6</td>
<td>0.6</td>
<td>0.79</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>Top 20%</td>
<td>0.59</td>
<td>0.83</td>
<td>0.76</td>
<td>0.51</td>
<td>0.78</td>
<td>0.34</td>
<td>0.91</td>
<td>0.64</td>
<td>0.86</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Bottom 20%</td>
<td>0.68</td>
<td>0.69</td>
<td>0.33</td>
<td>0.78</td>
<td>0.74</td>
<td>0.38</td>
<td>0.17</td>
<td>0.17</td>
<td>0.58</td>
<td>0.29</td>
<td>0.35</td>
</tr>
<tr>
<td>AFE</td>
<td>Complete</td>
<td>0.69</td>
<td>0.57</td>
<td>1.42</td>
<td>0.94</td>
<td>0.69</td>
<td>1.39</td>
<td>1.04</td>
<td>0.92</td>
<td>0.96</td>
<td>1.33</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Top 20%</td>
<td>1.3</td>
<td>0.57</td>
<td>0.88</td>
<td>1.24</td>
<td>1.67</td>
<td>1.68</td>
<td>1.11</td>
<td>0.8</td>
<td>1.07</td>
<td>1.34</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>Bottom 20%</td>
<td>0.63</td>
<td>0.53</td>
<td>2.28</td>
<td>1.07</td>
<td>0.54</td>
<td>0.86</td>
<td>1.1</td>
<td>1.16</td>
<td>0.9</td>
<td>1.07</td>
<td>0.99</td>
</tr>
</tbody>
</table>

ing the optimal transport problem to arrive at the set similarity coefficient – an analogy to the pairwise Tanimoto similarity. While we are not aware of other scientific literature utilizing the former in the same context, we argue that this is a simple, non-parametric way to describe the chemical similarity across two different distributions of compounds. A more qualitative approach is provided by embedding and visualizing chemical structures in two-dimensional space with the UMAP$^{22}$ algorithm. Here, insights can be obtained by analyzing the local structure in the embedding space, often combined with the coloring of points based on properties of interest. Of particular note is the use of publicly available data from ChEMBL$^{23,24}$ for the fitting. This allows us to create an independent reference space and embed the compounds of interest post fitting (inductively), enabling a consistent visualization and comparison of various data sets across different teams. Moreover, embedding compounds into a reference space could serve as a way of sharing chemical space distributions across the industry without revealing the actual chemical structure.

With regard to model development, a major challenge we had to address was the question of generalizability beyond a small training data set. To that end, neural network-based models offer an unparalleled level of flexibility when it comes to incorporating additional data. When training on the human PK data alone, no significant difference in performance was observed between tree-based and neural network-based models. The inclusion of related in vitro and non-clinical in vivo data into the training schedule ultimately delivered the best results with respect to prediction accuracy and correlation in cross-validation experiments. We found that the most beneficial mode of training is a set-up where one model is trained to predict both human PK and in vitro properties in a multi-task fashion. While we did not evaluate the predictive performance on the in vitro properties in this work, we note that the resulting model successfully integrates both data modalities in a meaningful way. In contrast, somewhat disappointing results were obtained...
for the two self-supervised pre-training strategies MolCLR and BRICS. In both cases, the predictive performance of the fine-tuned model was worse than that of the baseline, which was trained solely on the human PK data. Further investigation is needed to determine whether a more elaborate training schedule can alleviate this issue, or whether this is due to the self-supervised tasks being too distinct from the task of human PK prediction. Nevertheless, we see value in the newly-introduced task of BRICS classification as a self-supervised pre-training method and will evaluate its performance on other downstream tasks.

Overall test set performance of our predictions is similar to those reported in recent studies from Miljković et al. or Lombardo et al. However, communicating how to use our predictive models is key when making them available to drug design teams. We do not claim the same level of accuracy across all prospective compounds in question. Instead, we equip our models with an actionable measure of epistemic prediction uncertainty. Based on this output, users can assess whether their chemical space of interest is well-captured by the model and whether predictions should be trusted. In effect, we have shown how this translates to the ability to distinguish between well- and poorly-performing predictions. This functionality is essential for all applications where a constant level of model performance across the complete input space can not be guaranteed. Among the tested methods for uncertainty quantification, direct prediction of mean and variance assuming a Gaussian prior produced the most meaningful results. A further increase in both model accuracy and quality of the uncertainties could be observed when coupling the technique with ensembling. Considering the predictive performance of individual PK endpoints, we could predict clearance, bioavailability, time at maximum concentration, half times (IV & PO), volume of distribution and apparent volume of distribution with an AAFE < 3. Moreover, application of uncertainties allowed to identify subsets of data with increased performance where 8 out of 11 endpoints are predicted with an AAFE < 2.5 (namely, CL, C_{max}, F, T_{max}, T_{\text{half,IV}}, T_{\text{half,PO}}, V/F and V), proving that machine-learning predictions of clinical PK at zero experimental costs are highly feasible.

**Conclusion**

Our research has demonstrated the potential of computational models in predicting human pharmacokinetics (PK) at the drug design stage. We have curated one of the largest and most diverse datasets of clinical PK data, which has enabled us to develop models that predict key PK properties with significant accuracy. Our approach, which integrates in vitro and in vivo data and is capable of returning meaningful uncertainties, represents a substantial advancement in the field of PK modeling. We believe that our work is a significant step towards integrating PK predictions into early-stage drug design, potentially reducing the number of nonclinical studies and accelerating the development of new therapeutics. Despite the above, many more opportunities in the field of PK modeling are yet unexplored. For example, a closer look into self-supervised pre-training strategies or better methods for uncertainty quantification is warranted. Integration of additional data modalities as input features is another relevant topic of research. Finally, the exploration of explainable models for PK prediction is of interest, and a direction we would like to focus on in the future.

**Supporting Information Available**

PDF file with additional descriptors and of the human PK data set per source, tables containing all discussed metrics from the cross-validation experiments, additional descriptors and metrics for the test set predictions, and an overview of the GPS architecture and hyperparameters.

**Acknowledgement** L.K. acknowledges funding by the Roche Postdoctoral Fellowship (RPF) programme.
Author contributions statement: All authors participated in the study design and writing of the manuscript. L.K. participated in the data collection and analysis, as well as the development, implementation and evaluation of the ML models. N.M., K.G.Z., T.S. and L.S.-P. participated in the data collection and analysis, as well as the evaluation of the ML models. M.Z. participated in the development and evaluation of the ML models.

Competing interests: M.Z. is funded by Harvard University. All other authors are employees of F. Hoffmann-La Roche AG.

References


(6) Lombardo, F.; Bentzien, J.; Berellini, G.; Muegge, I. Prediction of Human Clearance Using In Silico Models with Reduced Bias. *Molecular Pharmaceutics* 2024,


(10) Manevski, N.; Umehara, K.; Parrott, N. Drug Design and Success of Prospective Mouse In Vitro–In Vivo Extrapolation (IVIVE) for Predictions of Plasma Clearance (CLp) from Hepatocyte Intrinsic Clearance (CLint). *Molecular Pharmaceutics* 2023,


