DeepCt: Predicting pharmacokinetic concentration-time curves and compartmental models from chemical structure using deep learning

Maximilian Beckers, Dimitar Yonchev, Sandrine Desrayaud, Grégori Gerebtzoff, and Raquel Rodríguez-Pérez*

Biomedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland

*Corresponding author

R.R.P. Phone: 41-795-42-2309, E-mail: raquel.rodriguez_perez@novartis.com

Abstract

After initial triaging using *in vitro* absorption, distribution, metabolism, and excretion (ADME) assays, pharmacokinetic (PK) studies are the first application of promising drug candidates in living mammals. Pre-clinical PK studies characterize the evolution of the compound's concentration over time, typically in rodents' blood or plasma. From this concentration-time (C-t) profiles, PK parameters such as total exposure or maximum concentration can be subsequently derived. An early estimation of compounds' PK offers the promise of reducing animal studies and cycle times by selecting and designing molecules with increased chances of success at the PK stage. Even though C-t curves are the major readout from a PK study, most machine learning-based prediction efforts have focused on the derived PK parameters instead of C-t profiles, likely due to the lack of approaches to model the underlying ADME mechanisms. Herein, a novel deep learning approach termed DeepCt is proposed for the prediction of C-t curves from the compound structure. Our methodology is based on the prediction of an underlying mechanistic compartmental PK model, which enables further simulations, and predictions of single- and multiple-dose C-t profiles.

Keywords: Machine learning, deep learning, mechanistic modeling, pharmacokinetics, compartmental analysis, concentration-time, time-exposure, drug discovery.

1. Introduction

Pharmacokinetic (PK) studies are an integral part of drug discovery to characterize drug's exposure, which is determined by the compound's absorption, distribution, metabolism, and excretion (ADME) processes [1], [2]. In contrast to PK studies, pharmacodynamics (PD) aims at understanding the drug's concentration-effect relationship. Hence, the combination of PK and PD (PK/PD) informs about the time course of drug response, which is key for dose predictions [3]. In PK studies, concentration-time (C-t) curves, also known as time-exposure curves, are obtained at various timepoints after compound's administration by a specific route and measured in a matrix, generally blood or plasma. To characterize C-t profiles, a variety of PK parameters can be derived, namely the area under the C-t curve (AUC), clearance (CL), half-life ($t_{1/2}$), maximum achieved concentration (C_{max}), or volume of distribution (V_d). Typically, PK behavior is first investigated in rodents after intravenous (i.v.) administration and might be followed by oral (p.o.) studies.

The mathematical modelling of C-t profiles has been traditionally based on a set of theoretical body 'compartments' through which the drug is distributed with linear kinetics [4]. Such compartmental PK models correspond to a system of ordinary differential equations (ODE), where a mass balance equation is defined for each compartment [5], [6]. Compartmental analyses have evolved towards physiologically based PK (PBPK) modeling[7], where the underlying ODE system is parameterized with physiological parameters, yielding more interpretable models. However, PBPK models are generally described by numerous compartments, leading to complex ODE systems (>100 equations). For predictive purposes, PBPK models are a method of choice with the integration of system- and compound-dependent parameters, including *in vitro* or *in vivo* data [8]. However, smaller compartmental models using up to three compartments are often sufficient to accurately model C-t curves from PK studies and provide important advantages such as the existence of analytical ODE solutions or more generalization ability due to the reduced set of parameters. Importantly, once the compartmental model is obtained for a compound, distinct administration scenarios can be simulated [6].

Machine learning (ML) advances have led to an increasing interest in ADME and PK predictions. Most ML efforts to predict compound properties have focused on *in vitro* ADME endpoints, such as intrinsic metabolic clearance and passive permeability, [9], [10], [11] and some approaches have also been put forward for an early ML-based prediction of *in vivo* PK parameters, either directly from chemical structure or from *in vitro* ADME data [12], [13], [14], [15], [16] [13], [17], [18], [19] [20], [21]. However, the direct ML-based prediction of C-t curves is an under investigated topic. To the best of our knowledge, the only ML approach for exposure-time curve prediction was recently proposed by AstraZeneca [13], but with limited success according to the authors. Their model predicted compound's concentration at specific timepoints using *in vitro* data as input, preventing its application at the drug design stage, and without mechanistically modelling the underlying ADME processes. A strategy to merge the benefits of PBPK and ML was recently proposed by colleagues at Bayer [14], [22]. The authors proposed a surrogate ML model consisting of a neural network to map the inputs of a PBPK model to its AUC and C_{max} readouts. The surrogate model circumvents the expensive computations of solving the large ODE system, but only estimates the derived PK parameters (AUC and C_{max}). To obtain the C-t profile model's output was mapped onto an actual PBPK model, [8] yielding considerable errors for both V_d and $t_{1/2}$ (mfce of 7.9 and 3, respectively).

Herein, DeepCt is proposed as a novel deep learning strategy for the prediction of C-t profiles from compound's structure and is applied to *in vivo* rat PK predictions. Our algorithm includes the direct prediction of PK compartmental models, leveraging the advantages of mechanistic and data-driven models. Specifically, a deep neural network is used for the prediction of a compartmental model's parametrization from which the full C-t profile can be inferred. As the underlying compartmental model is analytically solved, surrogate models are not required and the ML model enables further applications, such as investigation of multiple dosing regimens.

2. Results and Discussion

2.1. Principles of compartmental models' and C-t profiles' predictions

A deep learning approach to predict C-t curves solely from chemical structure was designed. In such algorithm, the input is a molecular representation and the constants of a compartmental model constitute the outputs. More details about the ML system are given in the following. Here, the model was applied to rat PK predictions, and an overview of the dataset used for modeling is shown in **Figure S1**. The dataset consisted of ~21,000 experiments, with ~13,000 and 9,000 from i.v. and p.o. administrations, respectively. Approximately 14,000 experiments were done using the blood as measurement matrix and ~ 8000 using plasma. Most studies were done in Sprague Dawley rats (~17,000) and Wistar Han rats (~5000) with a few experiments also done using Lewis and Brown Norway rats (<1000). For model building and evaluation, the data was split into three subsets according to the measurement date: (i) training, 80%; (ii) validation, 5%; and test (15%) sets. Therefore, the model was generated with the oldest studies, whereas the model's validation and testing was carried out with more recent data, which simulates prospective model's usage in pharmaceutical industry.

2.1.1. Prediction of compartmental models' constants

Figure 1a reports the schematic of a three-compartmental model, where the central compartment can be interpreted as the blood, the first peripheral compartment highly perfused tissues (e.g. muscles) and the second peripheral compartment as tissues barely perfused with blood (e.g. fatty tissues)[4] This compartmental system is mathematically described as an ODE system defined by a set of constants (see Section 4.3. Compartmental modeling). Specifically, the compartmental constants for i.v. administration are $CL_{iv}, V_{c,i.v.}, Q_{1,i.v.}, V_{1,i.v.}, Q_{2,i.v.}, V_{2,.i.v.}$ and for p.o. administration are $k_a, CL, V_c, Q_1, V_1, Q_2, V_2$. In cases where $Q_2 = 0$ and $Q_1 = Q_2 = 0$, two- and one-compartmental models' solutions are obtained, respectively.

As illustrated in **Figure 1b**, a deep learning algorithm was developed to predict the constants of a compartmental model from molecular structure. For that, molecules were numerically encoded using a data-driven representation known as MELLODDY embeddings (see 4.4. Molecular representation).

2.1.2. Prediction of C-t profiles

From the compartmental model constants, the full C-t profiles and, subsequently, the PK parameters defining such curves can be derived. The model was trained using a combined loss containing the mean absolute error (MAE) of the predicted C-t profiles in logarithmic scale and a penalty for distribution into higher compartments. Model's loss is given by

$$Loss = \frac{1}{n}\sum_{i\in 1,\ldots,n}\frac{1}{m}\sum_{j\in 1,\ldots,m}|\log\left(C_i\left(t_j\right)\right) - \log\left(C_i\left(t_j\right)_{pred}\right)| + w\frac{1}{4}(Q_{1,i.v.} + V_{1,i.v.} + Q_{1,p.o.} + V_{1,p.o.})$$

$$+ w\frac{1}{4}(Q_{2,i.v.} + V_{2,i.v.} + Q_{2,p.o.} + V_{2,p.o.}) \qquad (1).$$

Where *n* is the number of experiments in the respective batch used for model update and *m* the number of measured concentrations of experiment *i* at times t_j, $C_i\left(t_j\right)$. Moreover, *w* is an empirical constant that controls the strength of the penalty, which was set to 0.1. In the situation of using only a two- or one-compartmental model, $Q_{2,i.v.}$, $V_{2,i.v.}$, $Q_{2,p.o.}$, $V_{2,p.o.}$ and/or $Q_{1,i.v.}$, $V_{1,i.v.}$, $Q_{1,p.o.}$, $V_{1,p.o.}$ would be 0. Total exposures (AUCs), clearance (CL), $t_{1/2}$, V_d at steady state (V_{ss}), mean residence time (MRT), area under the first moment curve (AUMC) and bioavailability (F) were derived from the predicted C-t curves, as detailed in the Methods section.

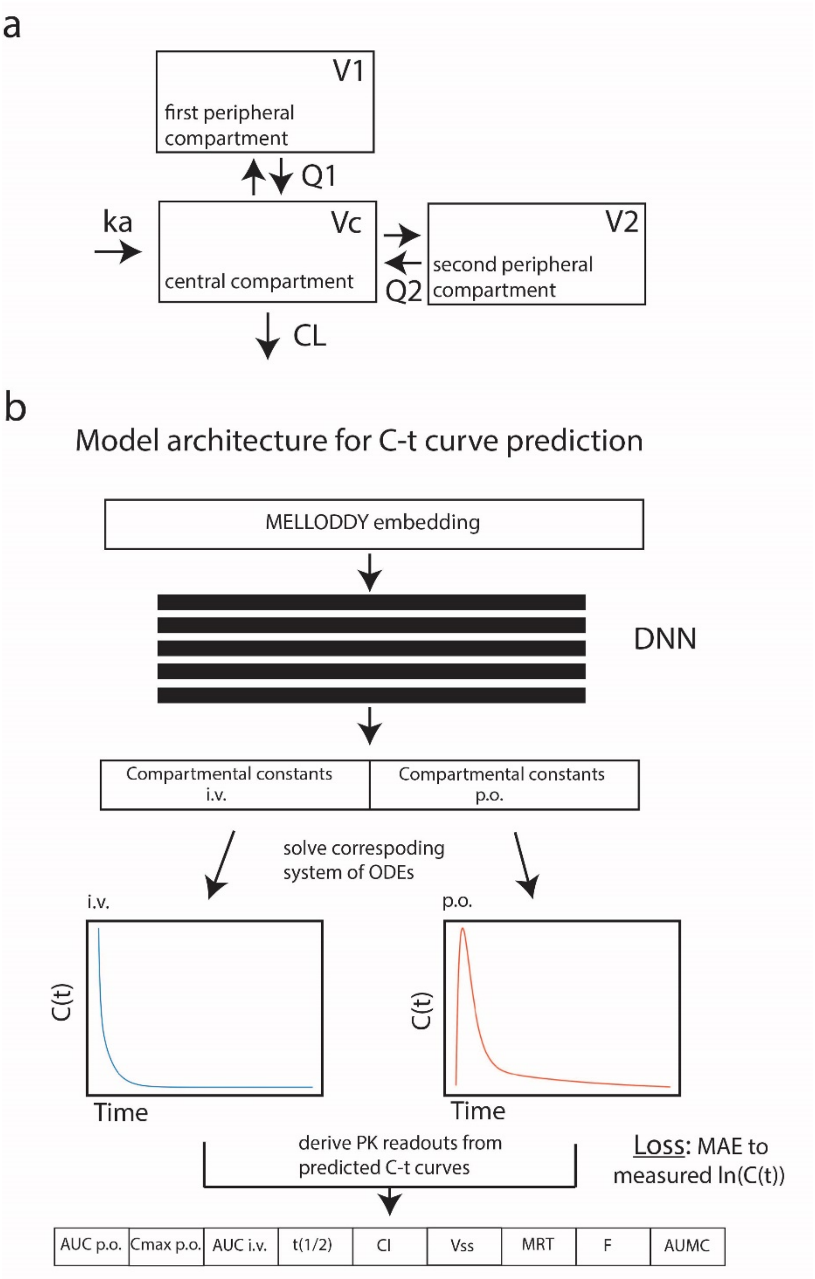

a
V1
first peripheral compartment
Q1
ka
Vc
central compartment
V2
second peripheral compartment
Q2
CL
b
Model architecture for C-t curve prediction
MELLODDY embedding
DNN
Compartmental constants i.v.
Compartmental constants p.o.
solve correspoding system of ODEs
i.v.
p.o.
C(t)
Time
C(t)
Time
derive PK readouts from predicted C-t curves
Loss: MAE to measured ln(C(t))
AUC p.o.
Cmax p.o.
AUC i.v.
t(1/2)
Cl
Vss
MRT
F
AUMC

Figure 1. ML algorithm for compartmental models and C-t curve prediction. Schematized are the (a) compartmental models, and (b) the deep neural network (DNN) used for prediction of compartmental constants with subsequent transformation to C-t curves and derivation of PK parameters.

2.2. Evaluating C-t curves' predictions

The quality of C-t profile predictions was first investigated for an underlying model with two compartments. A variety of metrics were calculated to compare experimental and predicted C-t profiles in the test set. Specifically, the coefficient of determination (R^2), Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ), and median fold change error (mfce) were calculated for each curve. To assess model's generalization ability, a temporal data splitting is often considered the gold-standard for property predictions in industry [23], [24]. Here, models were evaluated on the most recent experiments (15%). For comparison to recent work in literature [14], [22], models' performance was also evaluated on a randomized test set.

Table 1 reports the C-t curve prediction performance on the test sets, both with temporal and random compound data splits. For prospective applications, deep learning-based predictions showed a mfce ~2 for i.v. and ~2.8 for p.o. administration. Correlation coefficients approached 1 for the i.v. predictions, and were 0.77 (r) and 0.58 (ρ) for p.o. Moreover, R^2 illustrated predictive ability of the model for C-t curves after i.v. administration (0.66) but not for p.o. studies (-0.35). When evaluated on a random split, the model showed significantly improved performance measures, with a mfce of 2.45 compared to 2.81 reached with a temporal split (two-sided Brunner-Muzel test). Complete statistics including p-values for the differences are shown in **Table S1**.

Table 1. C-t curve prediction performance. Reported are performance metrics for the evaluation of C-t curve predictions, including the coefficient of determination (R^2), Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ), and median fold change error (mfce). Results are reported for the deep learning approach evaluated on time and random splits. Metrics considering replicated *in vivo* C-t measurements are also shown (exp. replicates).

Route of administration	**Strategy**	**R^2**	**r**	**ρ**	**mfce**
i.v.	Deep learning (time split)	0.66	0.97	0.99	2.03
	Deep learning (random split)	0.71	0.97	0.99	1.85
	Exp. replicates	0.96	0.99	1.00	1.20
p.o.	Deep learning (time split)	-0.35	0.77	0.58	2.81
	Deep learning (random split)	-0.18	0.79	0.64	2.45
	Exp. replicates	0.64	0.93	0.86	1.48

Since ML models' quality rely on the underlying training data quality, experimental uncertainty across study replicates was also characterized [9]. Importantly, the experimental variability analyzed herein refers to the intra-study variability (replicates), which is typically lower than the inter-study variability (different animals, but same protocol, day, etc.) [25]. Performance metrics were also calculated using replicated PK experiments and are reported in **Table 1**. Even though results show higher variability in p.o. curves, deep learning-based prediction errors were substantially larger than the experimental uncertainty coming from replicate measurements.

Figure 2 reports exemplary predictions of C-t profiles both for i.v. and p.o. scenarios, together with the measured data points (in red). These C-t curves correspond to prospective test compounds (temporal split). Visual inspection of the results indicate that i.v. curves' predictions more closely resemble the experimental PK curves than p.o. predictions, as anticipated from the performance metrics. Measurements' uncertainty is higher for p.o. compared to i.v. administration, indicating that the increased error of p.o. curve prediction might be due to the inherent ADME complexity of p.o. studies that lead to lower data quality.

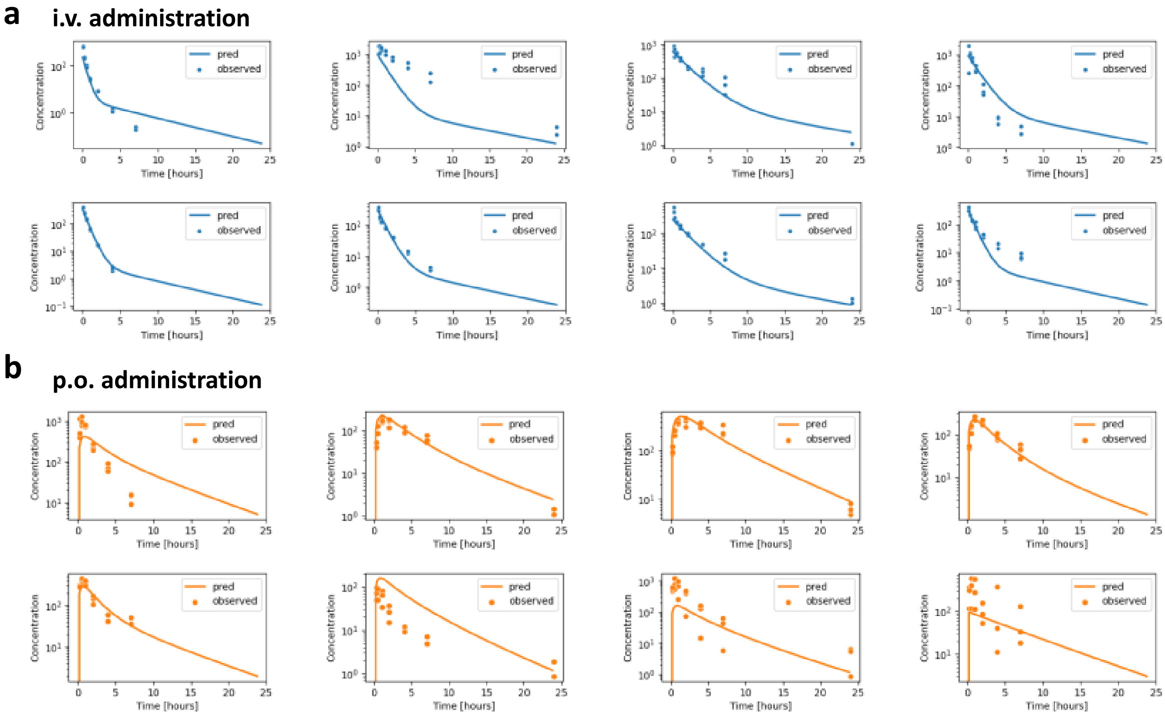
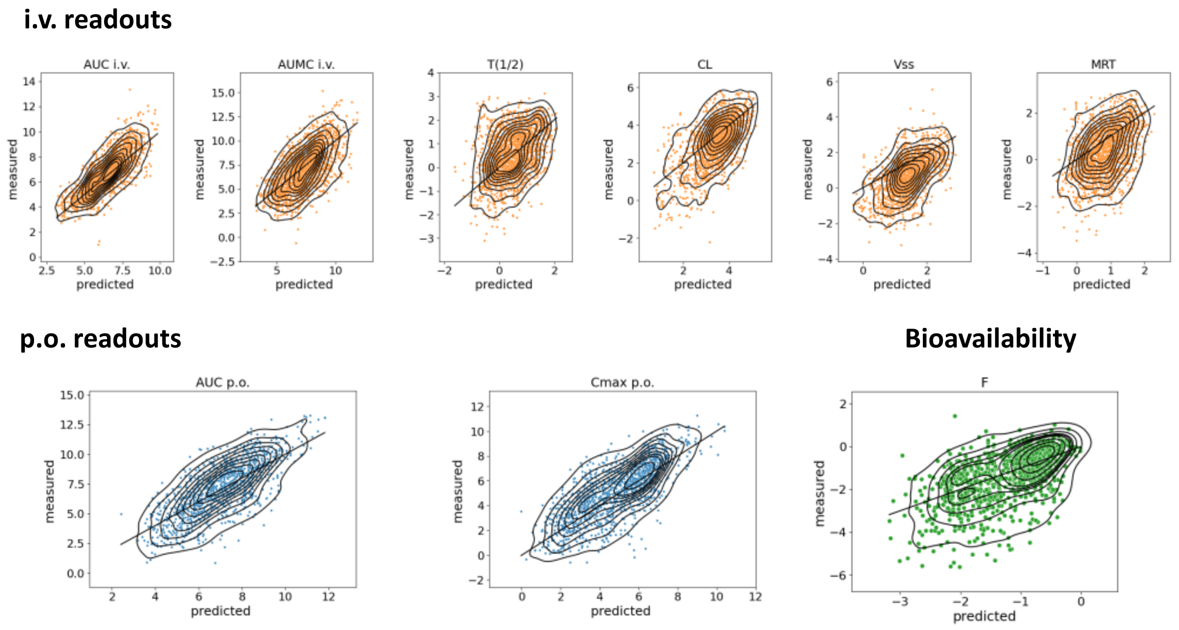


Figure 2. Predicted C-t curves. Predicted C-t profiles (solid lines) are reported together with the measured data points (dots) for exemplary compounds in the prospective test set (time split). The first two rows show predicted i.v. curves (blue) and the two bottom rows p.o. curves (orange).

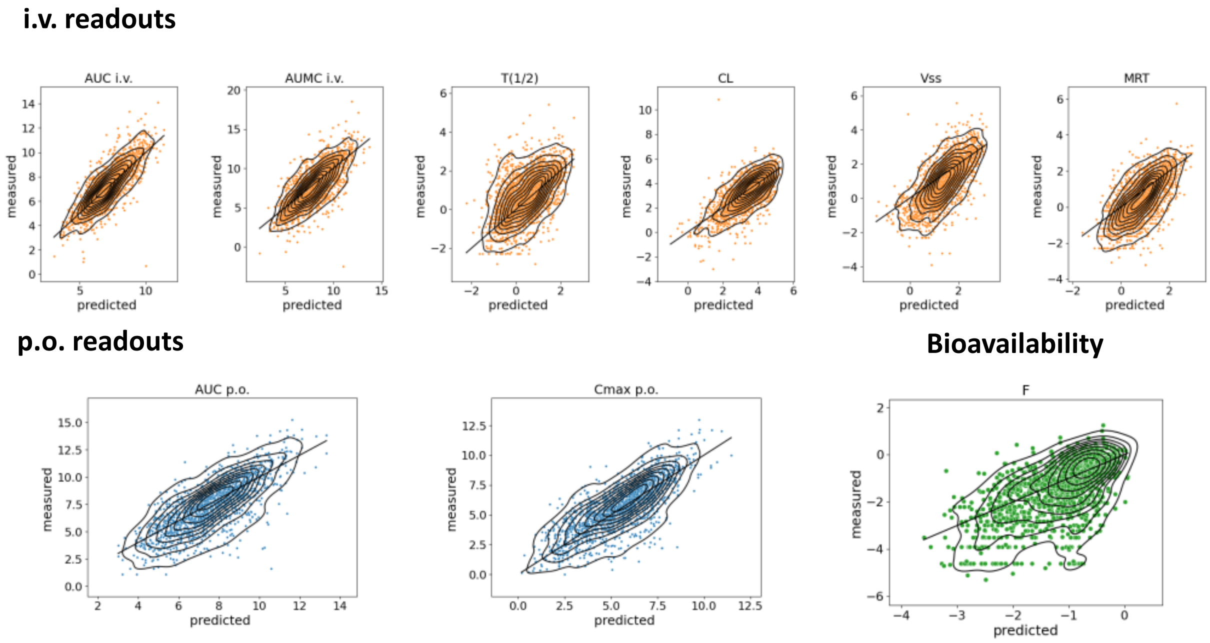
2.3. Evaluating PK parameters' predictions

In the context of drug discovery projects and to make decisions about compound prioritization, derived PK parameters are often used instead of the full C-t curves. In previous publications, fundamental PK parameters have constituted the prediction tasks [12], [14], [18], [21], [22]. To further assess the quality of C-t curve predictions and their usefulness for early PK assessments, key parameters were derived from the predicted time-exposure curves.

(A) Time split



(B) Random split



(C) Model performance with time and random train-test splits

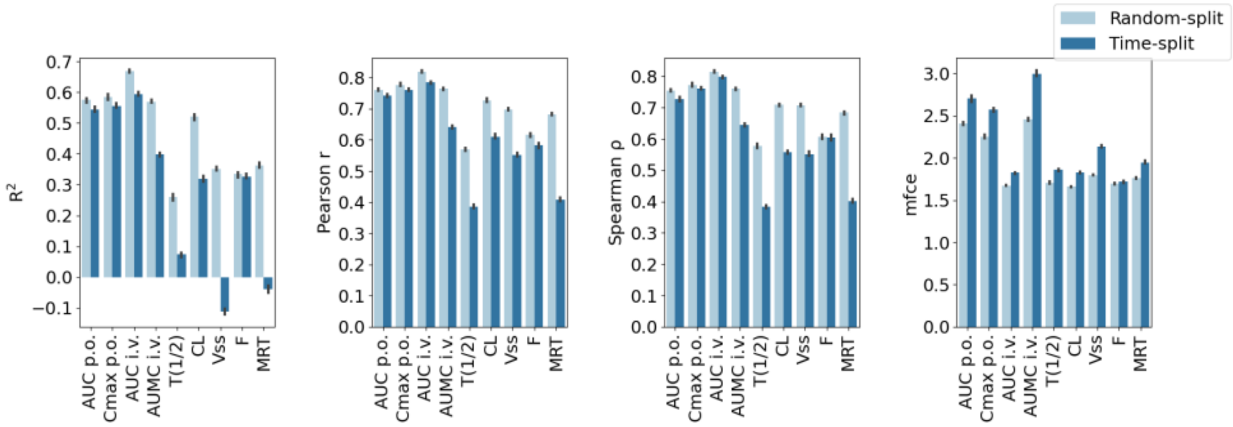


Figure 3. Prediction of derived PK parameters from C-t curves. Shown are predicted PK readouts derived from the predicted C-t curves vs. the measured values for evaluations with (A) time and (B) random data splits. Readouts after i.v. administration are shown in the top row (orange), whereas p.o. readouts (blue) and bioavailability (green) are shown in the bottom row. (C) Reported are performance metrics for the prediction of derived PK parameters with random (light blue) and time (blue) data splits. The median and 99% confidence intervals (error bars) obtained through bootstrapping are shown.

Figure 3 reports the comparisons of predicted versus observed PK readouts for models' evaluation of temporal and random data splits. Overall, endpoints from i.v. studies were estimated more accurately than the ones from p.o. studies, which is agreement with the increased accuracy of i.v. C-t curves' prediction compared to p.o. profiles. When evaluated prospectively (**Figure 3a and c**), model's mfce values ranged from 1.80 to 2 for most i.v.-related parameters. Furthermore, bioavailability was predicted with a mfce of 1.7, and AUC p.o. and Cmax p.o. predictions showed mfce of ~2.7.

Table 2. Derived PK parameter prediction performance. Reported are performance metrics for the evaluation of derived PK parameter predictions, including the coefficient of determination (R^2), Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ), and median fold change error (mfce). Results are reported for the deep learning approach evaluated on time and random splits.

Route of administration	**Parameter**	**R^2 (random/time-split)**	**r (random/time-split)**	**ρ (random/time-split)**	**mfce (random/time-split)**
i.v.	AUC	0.67/0.60	0.82/0.79	0.81/0.8	1.67/1.83
	AUMC	0.57/0.39	0.77/0.64	0.76/0.64	2.46/2.98
	CL	0.52/0.32	0.73/0.61	0.71/0.56	1.66/1.83
	Half-life	0.26/0.07	0.57/0.38	0.58/0.38	1.71/1.87
	Vss	0.35/-0.11	0.7/0.55	0.71/0.55	1.9/2.13
	MRT	0.37/-0.04	0.68/0.41	0.69/0.4	1.76/1.96
p.o.	AUC	0.57/0.54	0.76/0.74	0.75/0.73	2.4/2.68
	F	0.33/0.33	0.61/0.59	0.61/0.6	1.7/1.72
	Cmax	0.58/0.56	0.78/0.76	0.77/0.76	2.26/2.57

Figure 3b shows the predicted versus observed PK readouts for the proposed model for a random subset of compounds. As expected, evaluating the same prediction approach on a random split led to significantly improved predictions for all parameters. The results are summarized in Table 2, witchcomplete statistics shown in **Table S2**. Mfce values of ~1.6 were obtained for AUC i.v., ~2.4 for AUC p.o. and ~2.3 for C_{max} after p.o. administration (**Figure 3b and c**).

2.4. Comparison to the direct prediction of derived PK parameters

As a control, the performance of PK parameters' estimation from predicted C-t curves was compared to a direct ML-based prediction of rat PK parameters. A multi-task (MT) learning approach was used to predict four PK parameters, namely AUC i.v. and p.o, C_{max} p.o., t(1/2) i.v. and MRT. As shown in **Figure S2**, a muti-task deep neural network (MT-DNN) was trained using the same molecular representation as input (MELLODDY embeddings). The MT-DNN model was composed by some shared layers as well as task-specific layers and was generated using the same training data. More specifically, model has been trained and evaluated using the same temporal data splits as the C-t curve model**.**

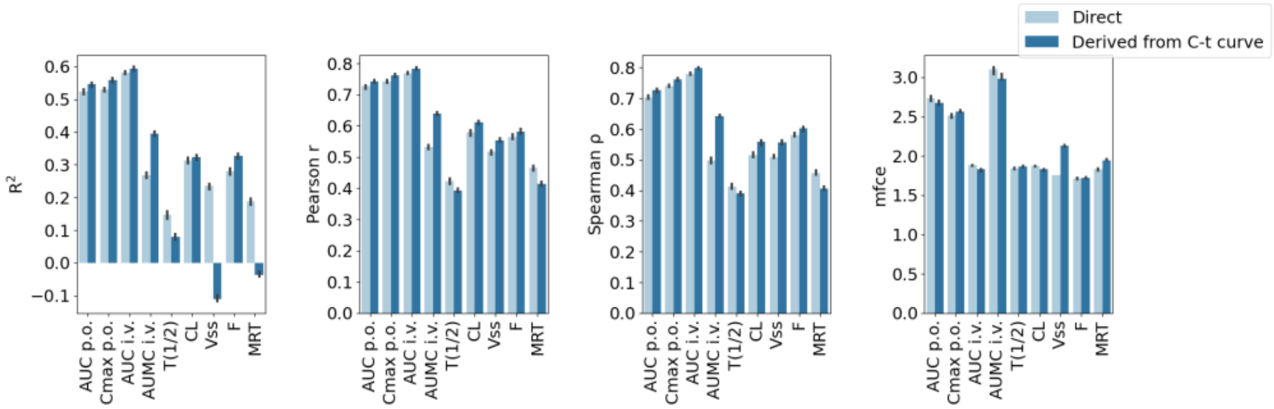


Figure 4. Benchmarking PK parameter prediction. Performance measures are shown for the different predicted PK parameters and compared between direct prediction (light blue) and derived from the predicted concentration time profile. The bars show the median and the error bars are 99% confidence intervals.

Figure 4 reports the comparison of both ML approaches in terms of the considered performance measures (complete statistics are shown in **Table S3**), and **Figure S3** shows the predicted versus observed PK parameters from the direct ML-based PK readouts' prediction. Overall, roughly equivalent performance was achieved with both models, showing that the prediction of complete C-t curves via prediction of a hidden compartmental model is not detrimental of predictive performance while it provides additional information. In contrast, for both Spearman's and Pearson's correlation coefficients predictions via C-t curves yielded significantly improved estimations of PK parameters for 7 out of the 9 considered parameters (only half-life and MRT were predicted slightly better with direct prediction).

2.5. Benchmarking compartmental models' complexity

In our approach for C-t curves' prediction, the complexity of the underlying compartmental model must be set (i.e. the number of distinct compartments). To this end, the model was implemented for up to three compartments and models' complexity was benchmarked. **Figure 5** compares prediction performance for one-, two-, and three-compartmental models. **Figure 5a** shows models' validation for C-t curve prediction, whereas **Figure 5b** reports the performance of predicted PK parameters (after derivation from predicted C-t profiles). When considering a two-compartmental model, the ML algorithm achieved promising performance for the prediction of C-t curves across all considered metrics. Reducing the number of compartments to one resulted in significantly worse predictions for C-t curves as well as most derived PK parameters ($p<0.01$, **Table S4**), with the median R^2 decreasing from ~0.7 to ~0.4 for i.v, C-t curves' prediction, indicating a tendency to underfitting. While a single compartment was usually not enough to model the distribution processes, three compartments did not result in significant differences compared to two compartments, neither for the prediction of C-t profiles nor derived PK parameters ($p<0.01$, **Tables S4, S5, S6** and **S7**).

Figures S4 shows the relationship between predictions and *in vivo* outcomes using one- and three-compartmental models. For PK parameters' prediction, results indicate that especially the half-lives are predicted differently. With a one-compartmental model there is no distribution phase and modelling it seemed necessary to improve the C-t curve predictions. Therefore, further evaluations were done with the consideration of a two-compartmental model.

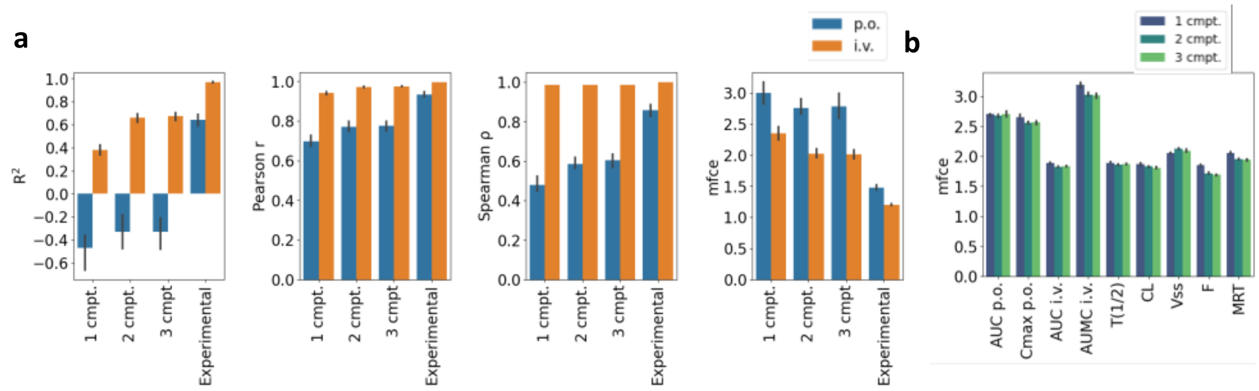


Figure 5. Benchmarking compartmental models' complexity. Performance evaluation for models with one, two or three compartments. (a) The goodness of individual curve predictions are shown with median R^2 , r, ρ and mfce metrics (from left to right). Results are reported for both i.v. (orange) and p.o. (blue) administration. (b) Shown are mfce values for the derived PK parameters obtained from predicted C-t profiles.

2.6. Benchmarking of model features

Different sets of input features for the ML model were benchmarked. First, experimental conditions were encoded as input features in addition to the MELLODDY embeddings to assess whether this information could further improve the model. Moreover, other molecular representations were tested to compare with MELLODDY embeddings. To this end, the DeepCt model was trained using Morgan fingerprints as compound representation together with structural descriptors (see Methods). Moreover, an additional DeepCt model was generated based on predicted ADME properties as input features. **Figure 6** reports the prediction performance for the C-t and PK parameters' prediction across the different feature sets.

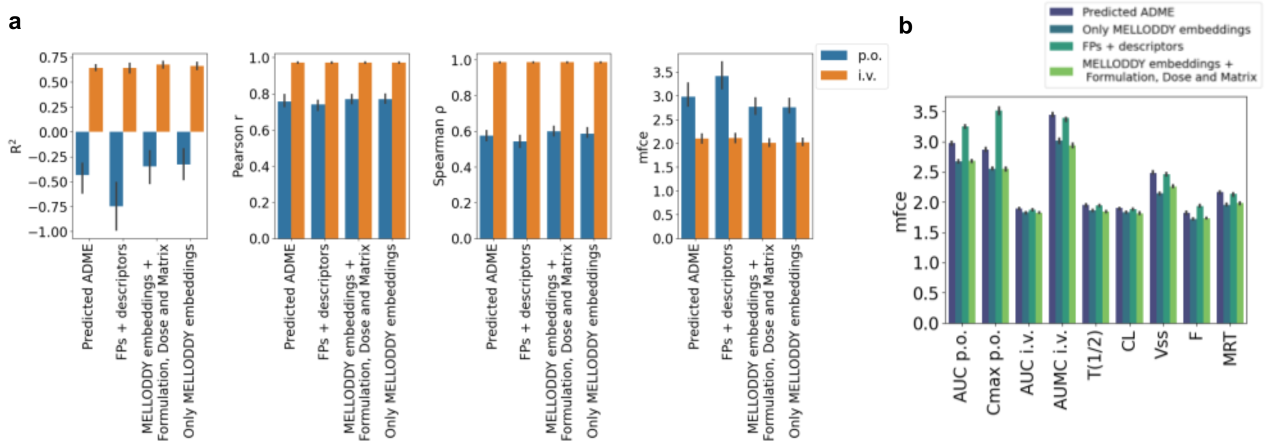


Figure 6. Benchmarking different sets of input features. Performance comparison for models based on different sets of input features (MELLODDY embeddings, MELLODDY embeddings with the addition of experimental conditions, Morgan fingerprints and 2D descriptors, and MELLODDY-based ADME predictions. (a) The goodness of individual curve predictions are shown with median R^2, r, ρ, and mfce metrics (from left to right). Results are reported for both i.v. (orange) and p.o. (blue) administration. (b) Reported are mfce values for the derived PK parameters obtained from predicted C-t profiles.

2.6.1. Encoding experimental conditions

The inclusion of experimental information as input was investigated. The dose, measured matrix (i.e. blood or plasma) and formulation (i.e. solution or suspension) were encoded as features. Results did not show evidence that such additional features improved predictions (**Figure 6**). Complete statistics are shown in **Table S8** for C-t curve predictions and **Table S9** for derived PK readouts. No statistically significant improvement could be found for the prediction of C-t curves ($p<0.01$, Brunner-Munzel test).

2.6.2. Alternative molecular representations

The effect of using different molecular representations as input to the model was also explored. Models were also generated based on (i) Morgan fingerprints of radius 2 and 1024 bits [26] concatenated with two-dimensional structural descriptors [27], and (ii) the predicted ADME property profile (from the MELLODDY model). To use the predicted ADME property profile, instead of calculating the model's embeddings, MELLODDY model's output predictions are used as features for our ML algorithm. For this we used 33 endpoints corresponding to *in vitro* ADMET (solubility assays, permeability assays, intrinsic clearance in different species, etc.). MELLODDY embeddings resulted in the best performance for both the prediction of C-t curves as well as the derived PK readouts (**Figure 6**). Statistically significant improvements of using the MELLODDY embeddings ($p<0.01$, two-sided Brunner-Munzel test) were found both compared to using the predicted ADME (**Table S10 and S11**) as input as well as the Morgan fingerprints (**Table S12 and S13**).

2.7. Prediction of C-t profiles for multiple dosing schemes

The deep learning-based estimation of the compartmental constants that describe C-t curves enabled the investigation of multiple dosing regimens. In such settings, the drug is repeatedly administered in order to achieve plasma concentrations above a minimum therapeutically active concentration level and below a toxic concentration level.

DeepCt generated with MELLODDY embeddings, and an underlying two-compartmental model was used. With the predicted compartmental model's parameters, an easy extension of single dosing to different multiple dosing regimens was explored. Subsequent administered dose was added at selected timepoints to the central compartments or to the gastrointestinal tract in the case of i.v. or p.o. administration, respectively. **Figure 6** shows C-t profiles over five days for two exemplary marketed non-Novartis compounds, namely Venetoclax (ABT-199, here Compound 1) and Rivaroxaban (BAY 59-7939, here Compound 2). These profiles were the result of simulating three dosing regimens with a dose of 1mg/kg every 6, 12 and 24 hours for both p.o. and i.v. administrations.). Compound 1 had a higher half-life than compound 2, and a 24h dosing interval already led to visible accumulation after multiple dosing, which is only visible for more frequent dosing for compound 2. This is in agreement with published data, with Rivaroxaban exhibiting a short half-life of 0.9h (0.89h predicted by DeepCt) in rats [28], while Venetoclax is known to exhibit longer half lifes of several hours (>10h in human [29], 3.66 predicted by DeepCt) . All in all, our simulations demonstrate that ML-based predictions of the underlying compartmental PK models enable the prediction even more complex readouts than single dose C-t curves.

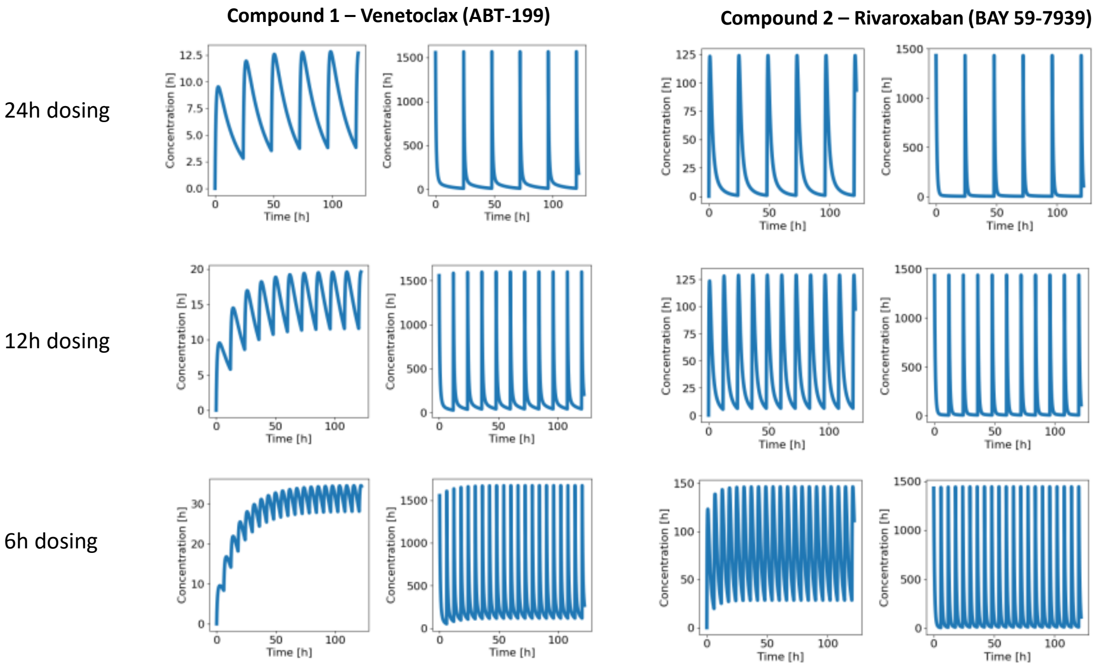


Figure 6. Prediction of C-t profiles for multiple dosing schemes. Predicted C-t curves are reported for two compounds after both i.v. and p.o. administration, respectively. Dosing every 24h hours (top row), every 12 hours (middle row) and every 6 hours (bottom row) is shown with a 1 mg/kg dose each.

3. Conclusions

In this work, a novel ML based approach was developed for the prediction of C-t curves and the underlying compartmental PK model constants. Our model only takes as input the molecular structure and predicts the parametrization of the corresponding PK compartmental model, which is a standard technique for modeling the ADME processes that govern the shape of C-t profiles. Given the compartmental parameters, C-t curves are obtained as the analytical solution of the respective ODE system.

Using the developed ML model, C-t profiles were predicted with reasonable accuracy, with median fold change errors of 2.0 and 2.8 after i.v. and p.o. administration, respectively. Moreover, results showed that PK parameters (e.g. AUC, F, etc.) derived from the predicted C-t curves exhibit similar accuracy to their direct PK prediction without modelling the underlying compartmental model. Hence, DeepCt keeps state-of-art performance while providing with valuable additional information, namely a mechanistic PK model that goes beyond the derivation of PK readouts and enables further simulations. Our work demonstrates first applications of our ML methodology beyond single dose studies and shows how the predicted compartmental model can be used to simulate the time profiles of multiple dosing schedules solely from chemical structures.

Some challenges remain and might be further considered in future work. For instance, as the absorption process adds additional complexity, C-t curves after p.o. administration showed larger experimental uncertainty and were also substantially harder to predict than i.v.,. Identifiability of the compartmental models might play an important role, as multiple solutions can explain the same or similar C-t curves. This effect is usually more pronounced for p.o. administration than for i.v. An additional limiting factor is the heterogeneity inherent to *in vivo* PK datasets, since data was generated at multiple laboratories, with different formulations, and strains, among others.

Taken together, the deep learning model DeepCt enables the prediction of PK C-t curves from chemical structures and is trained end-to-end using measured concentrations. Given that our model reaches state-of-art performance for PK parameters (e.g. AUC, bioavailability) but also predicts an underlying PK compartmental model and C-t profile, it can be envisioned that DeepCt and derivates of it will become a preferred approach for *in vivo* PK predictions.

4. Materials and Methods

4.1. Protocol of *in vivo* rat PK studies

The animal experiments were performed in accordance with the global Novartis Animal Welfare Policy and Standards and to the Animal Welfare legislation and regulations of the country where the study was conducted (USA or Switzerland). All rodents used here were obtained from approved vendors and were kept under standard conditions after their arrival at the test facilities. They were acclimatized for several days and were housed under controlled environmental conditions (optimal health conditions, 22 °C in a special, acclimatized pathogen-free animal room with 12h dark-light cycles) with ad libitum access to standard food and tap water before dosing and during the entire experimentation period.

The PK studies were conducted as discrete dosing or cassette dosing (i.e., mixture of 6 compounds dosed simultaneously). For rat PK studies, four to six days before drug administration, male Sprague Dawley rats (body weight 250-300 g) were anesthetized, and then, under aseptic conditions, two catheters were surgically implanted into the left and right jugular veins for drug administration and blood collection, respectively. The catheters were exteriorized at the neck. Animals received analgesic treatment before surgery and subsequently at appropriate times after surgery, and animals were kept individually in standard cages. After recovery, the cannulated rats were dosed either intravenously via the catheter at a dose of 1 mg/kg of drug substance solubilized in a mixture of N-1-methylpyrrolidone (NMP) and polyethylenglycol 200 (PEG200) with an administration volume of 0.5 mL/kg or orally by gastric gavage at a dose of 10 mg/kg as a suspension in a mixture of Methylcellulose and Tween 80 with an administration volume of 5 mL/kg. At different time points, over 24 h after dosing, blood (EDTA, ~10-20 µL) was collected via the other catheter. Immediately after collection, these whole blood samples were frozen on dry-ice and then stored at −20°C until LC/MS/MS analysis for parent drug determination. For bioanalytical investigation of blood samples, protein precipitation was performed by mixing an aliquot of blood with acetonitrile and centrifuged at 4 °C. The supernatant was transferred into a microtiter plate and an aliquot of each sample was injected into the LC-MS/MS system for analysis. Due to the utilization of a large data set, comprising many years of measurements at

Novartis, parts of the protocol might differ for some of the compounds measured in PK studies. For instance, some studies might have carried out in plasma or for Wistar rats.

Analysis of the experimental concentration curves was done using a non-compartmental approach, according to internal PK guidelines. The PK calculations were performed on individual concentration profiles. All calculations were based on the compounds' free form. Briefly, the apparent terminal slope λz (rate-constant in h-1) of the semilogarithmic concentration-time curve was estimated between at least the 3 last measured time point concentrations (with a square of correlation coefficient, also named goodness of fit statistic, $R^2 > 0.75$); then the apparent elimination half-life ($t_{1/2}$,z) was calculated as $t_{1/2z} = \ln 2 / \lambda z$. The areas under the curve (AUC) were calculated by the linear trapezoidal rule for increase and logarithmic for decrease and extrapolated to infinite time as AUC = AUC_{last} + Clast/λz, where AUC_{last} is the area under the curve between zero and the last measurable time point (t_{last}) and C_{last} the last measurable blood concentration (i.e. last level data above the limit of quantification). The extrapolation of the AUC from the last time point to infinite (i.e. $C_{last}/\lambda z$) did not exceed 25% of the AUC_{inf}. Systemic plasma clearance (CL) data were calculated as: CL = $Dose_{iv}/AUC_{iv}$. The oral absolute bioavailability (%) was estimated using the dose-normalized AUC oral/intravenous percentage. The maximal blood concentration observed after oral administration corresponds to C_{max}.

4.2. Dataset preparation and description

Rat i.v. and p.o. studies in blood or plasma were considered for modeling. For cases with blood/plasma partition coefficients available, corrections were applied to transform plasma to blood concentrations. Dosage forms were solution or suspension, and only single dose studies were kept. Concentrations below quantification limit were considered as missing values. Instances with missing concentrations were also included in the dataset since they might be due to compounds with fast clearance. Experiments were discarded for animals with less than two concentration values measured or a unique concentration value in all the experiment.

The final dataset contained ca. 21000 rat experiments, of which ca. 8000 were i.v. and 13000 p.o. administrations (**Supporting Figure 1**).

The data set was split into training (80%), validation (5%), and test (15%) sets, depending on the study report date. The training set was used for model generation, the validation set for early estopping (as further detailed in Section 4.4.) and the test set for prospective performance evaluation.

4.3. Compartmental modeling

Compartmental analysis is commonly used in PK to model the kinetic behavior of a compound in the body after administration [30], [31]. A three-compartmental model (**Figure 1a**) is defined by absorption (in case of p.o. administration, modelled by the rate constant k_a), distribution from the central compartment to both the first and second peripheral compartment (modelled by intercompartmental clearances Q_1 and Q_2 and compartmental volumes V_1 and V_2, respectively) and elimination from the central compartment only (modelled by CL). Here, distribution to the peripheral compartments only happens from the central compartment. Using linear kinetics and for a p.o. bolus, this compartmental model results in the following ODE system for the concentration $C_G(t)$ in the gut, $C_c(t)$ in the blood, $C_1(t)$ in the first peripheral compartment and $C_2(t)$ in the second peripheral compartment:

$$\frac{dC_G(t)}{dt} = -k_a \cdot C_G(t) \quad (2)$$

$$\frac{dC_c(t)}{dt} = k_a \cdot C_G(t) \; - \; \frac{CL}{V_c} C_c(t) - \frac{Q_1}{V_c} C_c(t) + \frac{Q_1}{V_1} C_1(t) - \frac{Q_2}{V_c} C_c(t) + \frac{Q_2}{V_2} C_2(t) \quad (3)$$

$$\frac{dC_1(t)}{dt} = \frac{Q_1}{V_c} C_c(t) - \frac{Q_1}{V_1} C_1(t) \quad (4)$$

$$\frac{dC_2(t)}{dt} = \frac{Q_2}{V_c} C_c(t) - \frac{Q_2}{V_2} C_2(t) \quad (5)$$

This system can be solved by initializing $C_G(0)$ using the loading dose, and setting $C_1(0) = 0$ as well as $C_2(0) = 0$. In case of an i.v. bolus, k_a is set to 0 and $C_c(0)$ defined by the dose.

This initial value problem is then solved by the following function.

Let

$$a_0 = CL \cdot Q_2 \cdot \frac{Q_3}{V_1 V_2 V_3} \qquad (6)$$

$$a_1 = \frac{CL \cdot Q_3}{V_1 V_3} + \frac{Q_2 \cdot Q_3}{V_2 V_3} + \frac{Q_2 \cdot Q_3}{V_2 V_1} + \frac{Q_2 \cdot CL}{V_2 V_1} + \frac{Q_2 \cdot Q_3}{V_1 V_3} \qquad (7)$$

$$a_2 = \frac{CL}{V_1} + \frac{Q_2}{V_1} + \frac{Q_3}{V_1} + \frac{Q_2}{V_2} + \frac{Q_3}{V_3} \qquad (8)$$

$$p = a_1 - \frac{a_2^3}{3} \qquad (9)$$

$$q = \frac{2a_2^3}{27} - \frac{a_1 a_2}{3} + a_0 \qquad (10)$$

$$r_1 = \sqrt{-\frac{p^3}{27}} \qquad (11)$$

$$r_2 = 2r_1^{\frac{1}{3}} \qquad (12)$$

$$\lambda = \frac{1}{3}\arccos\left(-\frac{q}{2r_1}\right) \qquad (13)$$

$$\alpha = -\left(\cos(\lambda)\, r_2 - \frac{a_2}{3}\right) \qquad (14)$$

$$\beta = -\left(\cos\left(\lambda + \frac{2\pi}{3}\right) r_2 - \frac{a_2}{3}\right) \quad (15)$$

$$\gamma = -\left(\cos\left(\lambda + \frac{4\pi}{3}\right) r_2 - \frac{a_2}{3}\right) \quad (16)$$

The final concentration over time $C_c(t)$ in the central compartment, i.e. the blood, is defined as:

$$C_c(t) = D\left(A\, e^{-\alpha t} + B\, e^{-\beta t} + C\, e^{-\gamma t}\right) \quad (17)$$

Where *D* is the dose and, in the case of an i.v. bolus, *A, B* and *C* are:

$$A = \frac{\left(\frac{Q_1}{V_1} - \alpha\right)\left(\frac{Q_2}{V_2} - \alpha\right)}{V_c(\alpha - \beta)(\alpha - \gamma)} \quad (18)$$

$$B = \frac{\left(\frac{Q_1}{V_1} - \beta\right)\left(\frac{Q_2}{V_2} - \beta\right)}{V_c(\beta - \alpha)(\beta - \gamma)} \quad (19)$$

$$C = \frac{\left(\frac{Q_1}{V_1} - \gamma\right)\left(\frac{Q_2}{V_2} - \gamma\right)}{V_c(\gamma - \alpha)(\gamma - \beta)} \quad (20)$$

In the case of p.o. administration *A*, *B* and *C* are given by the following equations:

$$A = \frac{k_a\left(\frac{Q_1}{V_1} - \alpha\right)\left(\frac{Q_2}{V_2} - \alpha\right)}{V_c(k_a - \alpha)(\alpha - \beta)(\alpha - \gamma)} \quad (21)$$

$$B = \frac{k_a\left(\frac{Q_1}{V_1} - \beta\right)\left(\frac{Q_2}{V_2} - \beta\right)}{V_c(k_a - \beta)(\beta - \alpha)(\beta - \gamma)} \quad (22)$$

$$C = \frac{k_a\left(\frac{Q_1}{V_1} - \gamma\right)\left(\frac{Q_2}{V_2} - \gamma\right)}{V_c(k_a - \gamma)(\gamma - \alpha)(\gamma - \beta)} \quad (23)$$

These equations solve the ODE system for a compartment model with three compartments. When $Q_2 = 0$, the solution for a two-compartmental model is obtained, whereas $Q_1 = 0$ and $Q_2 = 0$ refers to the one-compartmental model's solution.

4.4. Non-compartmental analysis

Non-compartmental analysis is used to derive PK parameters from the C-t curves [32].Total exposures over time for p.o. administration, $AUC_{p.o.}$, and for i.v., $AUC_{i.v.}$ are calculated from the curves by the trapezoidal rule. Similarly, the $AUMC_{i.v.}$, i.e. the area under the first moment curve, is calculated using the trapezoidal rule from the first moment curve , which is given by the concentration times time C(t)*t. From the total exposures $AUC_{i.v.}$, $AUC_{p.o.}$ and corresponding doses $D_{i.v.}$ and $D_{p.o.}$, biovailability F is then calculated by

$$F = \frac{D_{i.v.}AUC_{p.o.}}{D_{p.o.}AUC_{i.v.}} \quad (24).$$

Clearance (CL) is defined as

$$CL = \frac{D_{i.v.}}{AUC_{i.v.}} \quad (25).$$

Half-life *t1/2* is calculated from the tail of the C-t curves by

$$t_{1/2} = \frac{\ln(2)}{\frac{C_{last}}{AUC_{i.v.} - AUC_{i.v.,last}}} \quad (26),$$

where $AUC_{i.v., \, last}$ is the AUC from time t=0 to the time where the concentration fell to 10% of its maximum, which we done here by C_{last}.

Mean residence time (MRT) is calculated by

$$MRT = \frac{AUMC_{i.v.}}{AUC_{i.v.}} \quad (27)$$

Moreover, the volume of distribution at steady state, V_{ss}, is then calculated by

$$V_{ss} = MRT * CL \qquad (28).$$

4.5. Deep learning model architectures and training

Two deep learning models were generated. The principal model is to predict compartmental models' constants and, subsequently, the C-t profile (DeepCt). The second model was generated as a control benchmark for the prediction of the derived PK parameters from molecular structure.

For both deep learning strategies, AdamW optimizer [33], early stopping [34], and cyclical learning rates [35] were applied. For early stopping, training was stopped after 10 epochs without further improvements on the validation set, and the best model was selected. For the cyclical learning rate scheduling, a base learning rate of 0.0001, a maximum learning rate of 0.01, and the "triangular2" cycling mode were chosen, and the number of iterations of one cycle was set to four times the batch size. A weight decay of 0.01 and a batch size of 32 were used. The final models are always an ensemble of 10 replicates, trained with different random initializations. Models are implemented in Python using *PyTorch* [36], *scikit-learn* [37], *scipy* [38] and *numpy* [39] libraries.

4.5.1. Compartmental models and C-t prediction

The models used for C-t profile predictions have five fully connected layers with 256 hidden units. Each layer is preceded by a dropout layer with a dropout rate of 0.3 and followed by batch normalization and ReLU activations. The loss function was the mean absolute error of the natural logarithm of the predicted and measured concentrations, respectively, together with an additional penalty on contributions of higher compartments (see section 2.1.2 for details).

4.5.2. Derived PK parameters

Models for derived PK readout prediction had three fully connected shared layers and three additional fully connected layers for each task. Each layer had 256 hidden units. The prediction tasks were AUC for both i.v. and p.o routes, C_{max} for p.o., and t(1/2) for i.v., which constituted the four output units of the network. The parameters Vss, CL and F were subsequently derived from the beforementioned predictions. Dropout layers (with a rate of 0.3), batch normalization, and ReLU activations were also considered. For this model, the loss function used was the mean absolute log error.

4.6. Molecular representation

Molecules were represented using the embeddings of a cross-pharma federated learning model termed MELLODDY. In the MELLODDY project, ten pharmaceutical companies built multi-task ML models through federated learning on a platform audited for privacy and security. Training data consisted of 2.6+ billion confidential activity data points, 21+ million physical small molecules, and 40+ thousand assays in on-target and secondary PD and PK[40] . Herein, the encodings from the last layer for a regression MELLODDY model were used as molecular representation and fed into the deep learning models for PK predictions.

For comparison, other molecular representations were tested. To this end a model using Morgan 2 fingerprints (bit-based, 1024 bits) [26] as well as structural descriptors calculated using RDKit [27], as described in Section 2.6.2.

4.7. Performance metrics

The fold change is a major criterion for assessing the performance of PK predictions [18], [41], [42], which can be aggregated for multiple predictions as the median fold change error *mfce* and is given by

$$\text{mfce} = \exp\left(\text{median}(\log(x) - \log(x_{pred}))\right) \qquad (29).$$

Other error measures we used are R^2, Pearson and Spearman correlations as well as the root mean squared error (RMSE).

4.8. Statistical analysis

Statistical significance of the differences of model performance was assessed by bootstrapping the test set and calculating p-values using the Brunner-Munzel test [43]. To account for the multiple testing problem originating from testing different readouts and performance measures simultaneously, p-values were further corretec for the FWER using Holm's approach [44].

References

[1] P. Bonate, *Pharmacokinetic-pharmacodynamic modeling and simulation*, vol. 9780387271972. 2006. doi: 10.1007/b138744.

[2] A. Ruiz-Garcia, M. Bermejo, A. Moss, and V. G. Casabo, ‘Pharmacokinetics in drug discovery’, *Journal of Pharmaceutical Sciences*, vol. 97, no. 2. 2008. doi: 10.1002/jps.21009.

[3] T. S. Maurer, D. Smith, K. Beaumont, and L. Di, ‘Dose Predictions for Drug Design’, *J Med Chem*, vol. 63, no. 12, pp. 6423–6435, 2020, doi: 10.1021/acs.jmedchem.9b01365.

[4] M. Holz and A. Fahr, ‘Compartment modeling’, *Adv Drug Deliv Rev*, vol. 48, no. 2–3, 2001, doi: 10.1016/S0169-409X(01)00118-1.

[5] A. Talevi and C. L. Bellera, ‘Two-Compartment Pharmacokinetic Model’, in *The ADME Encyclopedia*, 2022. doi: 10.1007/978-3-030-84860-6_59.

[6] A. Talevi and C. L. Bellera, ‘One-Compartment Pharmacokinetic Model’, in *The ADME Encyclopedia*, 2021. doi: 10.1007/978-3-030-51519-5_58-1.

[7] X. Zhuang and C. Lu, ‘PBPK modeling and simulation in drug research and development’, *Acta Pharmaceutica Sinica B*, vol. 6, no. 5. 2016. doi: 10.1016/j.apsb.2016.04.004.

[8] D. Naga, N. Parrott, G. F. Ecker, and A. Olivares-Morales, ‘Evaluation of the Success of High-Throughput Physiologically Based Pharmacokinetic (HT-PBPK) Modeling Predictions to Inform Early Drug Discovery’, *Mol Pharm*, 2022, doi: 10.1021/acs.molpharmaceut.2c00040.

[9] R. Rodríguez-Pérez, M. Trunzer, N. Schneider, B. Faller, and G. Gerebtzoff, ‘Multispecies Machine Learning Predictions of in Vitro Intrinsic Clearance with

Uncertainty Quantification Analyses', *Mol Pharm*, vol. 20, no. 1, pp. 383–394, 2023, doi: 10.1021/acs.molpharmaceut.2c00680.

[10] A. H. Göller *et al.*, 'Bayer's in silico ADMET platform: a journey of machine learning over the past two decades', *Drug Discov Today*, vol. 25, no. 9, pp. 1702–1709, 2020, doi: 10.1016/j.drudis.2020.07.001.

[11] S. Aleksić, D. Seeliger, and J. B. Brown, 'ADMET Predictability at Boehringer Ingelheim: State-of-the-Art, and Do Bigger Datasets or Algorithms Make a Difference?', *Mol Inform*, vol. 41, no. 2, p. 2100113, 2022, doi: 10.1002/minf.202100113.

[12] R. Stoyanova *et al.*, 'Computational Predictions of Nonclinical Pharmacokinetics at the Drug Design Stage', *J Chem Inf Model*, vol. 63, no. 2, pp. 442–458, 2023, doi: 10.1021/acs.jcim.2c01134.

[13] O. Obrezanova *et al.*, 'Prediction of in Vivo Pharmacokinetic Parameters and Time-Exposure Curves in Rats Using Machine Learning from the Chemical Structure', *Mol Pharm*, vol. 19, no. 5, pp. 1488–1504, 2022, doi: 10.1021/acs.molpharmaceut.2c00027.

[14] S. Schneckener *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*, vol. 59, no. 11, pp. 4893–4905, 2019, doi: 10.1021/acs.jcim.9b00460.

[15] M. Beckers, N. Sturm, N. Fechner, and N. Stiefl, 'Prediction of small molecule developability using large-scale in silico ADMET models', *J. Med. Chem*, vol. 66, no. 22, pp. 14047–14060, 2023.

[16] O. Obrezanova, ‘Artificial intelligence for compound pharmacokinetics prediction’, *Current Opinion in Structural Biology*, vol. 79. 2023. doi: 10.1016/j.sbi.2023.102546.

[17] Y. Kosugi and N. Hosea, ‘Prediction of Oral Pharmacokinetics Using a Combination of in Silico Descriptors and in Vitro ADME Properties’, *Mol Pharm*, vol. 18, no. 3, 2021, doi: 10.1021/acs.molpharmaceut.0c01009.

[18] G. Berellini, N. J. Waters, and F. Lombardo, ‘In silico prediction of total human plasma clearance’, *J Chem Inf Model*, vol. 52, no. 8, 2012, doi: 10.1021/ci300155y.

[19] G. Berellini, C. Springer, N. J. Waters, and F. Lombardo, ‘In silico prediction of volume of distribution in human using linear and nonlinear models on a 669 compound data set’, *J Med Chem*, vol. 52, no. 14, 2009, doi: 10.1021/jm9004658.

[20] A. Gruber, F. Führer, S. Menz, H. Diedam, A. H. Göller, and S. Schneckener, ‘Prediction of Human Pharmacokinetics From Chemical Structure: Combining Mechanistic Modeling with Machine Learning’, *J Pharm Sci*, Oct. 2023, doi: 10.1016/j.xphs.2023.10.035.

[21] F. Miljković *et al.*, ‘Machine Learning Models for Human in Vivo Pharmacokinetic Parameters with In-House Validation’, *Mol Pharm*, vol. 18, no. 12, 2021, doi: 10.1021/acs.molpharmaceut.1c00718.

[22] F. Führer, A. Gruber, H. Diedam, A. H. Göller, S. Menz, and S. Schneckener, ‘A deep neural network: mechanistic hybrid model to predict pharmacokinetics in rat’, *J Comput Aided Mol Des*, vol. 38, no. 1, p. 7, 2024, doi: 10.1007/s10822-023-00547-9.

[23] R. P. Sheridan, ‘Time-split cross-validation as a method for estimating the goodness of prospective prediction.’, *J Chem Inf Model*, vol. 53, no. 4, pp. 783–790, 2013, doi: 10.1021/ci400084k.

[24] G. A. Landrum, M. Beckers, J. Lanini, N. Schneider, N. Stiefl, and S. Riniker, ‘SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches’, *J Cheminform*, vol. 15, no. 1, p. 119, Dec. 2023, doi: 10.1186/s13321-023-00787-9.

[25] S. Winiwarter *et al.*, ‘Time dependent analysis of assay comparability: A novel approach to understand intra- and inter-site variability over time’, *J Comput Aided Mol Des*, vol. 29, no. 9, 2015, doi: 10.1007/s10822-015-9836-5.

[26] D. Rogers and M. Hahn, ‘Extended-connectivity fingerprints’, *J Chem Inf Model*, vol. 50, no. 5, pp. 742–754, 2010, doi: 10.1021/ci100050t.

[27] G. A. R. Landrum, ‘RDKit: Open-Source Cheminformatics Software’. 2016. [Online]. Available: http://www.rdkit.org/

[28] S. Roehrig *et al.*, ‘Discovery of the novel antithrombotic agent 5-chloro-N-({(5S)-2-oxo-3-[4- (3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl}methyl)thiophene-2-carboxamide (BAY 59-7939): An oral, direct factor Xa inhibitor’, *J Med Chem*, vol. 48, no. 19, 2005, doi: 10.1021/jm050101d.

[29] A. H. Salem, S. K. Agarwal, M. Dunbar, S. L. H. Enschede, R. A. Humerickhouse, and S. L. Wong, ‘Pharmacokinetics of Venetoclax, a Novel BCL-2 Inhibitor, in Patients With Relapsed or Refractory Chronic Lymphocytic Leukemia or Non-Hodgkin Lymphoma’, *J Clin Pharmacol*, vol. 57, no. 4, 2017, doi: 10.1002/jcph.821.

[30] F. Chaubet, V. Rodriguez-Ruiz, M. Boissière, and D. Velasquez, 'Pharmacology: Drug Delivery', *Encyclopedia of Biomedical Engineering*, pp. 440–453, 2019, doi: https://doi.org/10.1016/B978-0-12-801238-3.11007-4.

[31] S. A. Saghir and R. A. Ansari, 'Pharmacokinetics', *Reference Module in Biomedical Sciences*, 2018, doi: https://doi.org/10.1016/B978-0-12-801238-3.62154-2.

[32] J. Gabrielsson and D. Weiner, 'Non-compartmental Analysis', *Computational Toxicology: Volume I*, pp. 377–389, 2012, doi: 10.1007/978-1-62703-050-2_16.

[33] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization', *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[34] A. C. Ian Goodfellow, Yoshua Bengio, *Deep Learning - Ian Goodfellow, Yoshua Bengio, Aaron Courville - Google Books*. 2016.

[35] L. N. Smith, 'Cyclical learning rates for training neural networks', *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, 2017, doi: 10.1109/WACV.2017.58.

[36] A. Paszke *et al.*, 'PyTorch: An imperative style, high-performance deep learning library', *Adv Neural Inf Process Syst*, pp. 8024–8035, 2019.

[37] F. Pedregosa *et al.*, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, vol. 12, pp. 2825– 2830, 2011.

[38] T. E. Oliphant, 'SciPy: Open source scientific tools for Python', *Comput Sci Eng*, vol. 9, pp. 10–20, 2007, doi: 10.1109/MCSE.2007.58.

[39] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, 'The NumPy array: A structure for efficient numerical computation', *Comput Sci Eng*, vol. 13, no. 2, pp. 22–30, 2011, doi: 10.1109/MCSE.2011.37.

[40] W. Heyndrickx *et al.*, 'MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information', *J Chem Inf Model*, 2023, doi: 10.1021/acs.jcim.3c00799.

[41] L. Z. Benet and J. K. Sodhi, 'Investigating the Theoretical Basis for In Vitro–In Vivo Extrapolation (IVIVE) in Predicting Drug Metabolic Clearance and Proposing Future Experimental Pathways', *AAPS Journal*, vol. 22, no. 5. 2020. doi: 10.1208/s12248-020-00501-9.

[42] L. Z. Benet and J. K. Sodhi, 'Can In Vitro–In Vivo Extrapolation Be Successful? Recognizing the Incorrect Clearance Assumptions', *Clinical Pharmacology and Therapeutics*, vol. 111, no. 5. 2022. doi: 10.1002/cpt.2482.

[43] E. Brunner and U. Munzel, 'The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation', *Biometrical Journal*, vol. 42, pp. 17–25, 2000, doi: 10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U.

[44] S. Holm, 'A simple sequentially rejective multiple test procedure', *Scandinavian journal of statistics*, vol. 6, no. 2, pp. 65–70, 1979.

Acknowledgement

The authors thank Nikolas Fechner, Nikolaus Stiefl, Finton Sirockin, Allison Claas, Jan-Frederik Schlender, and Peter Ashcroft for helpful scientific discussions, and all present and former Novartis colleagues who contributed to generate and give access to the data that enabled these investigations. Authors also thank current and former Novartis colleagues involved in the data generation that has enabled this investigation. The authors would like to thank the MELLODDY consortium, especially Novartis colleagues (Ansgar Schuffenhauer, Nikolas Fechner, Nicholas Holway, and Noé Sturm) since model's embeddings were used herein. Maximilian Beckers acknowledges financial support from a Novartis Discovery Postdoctoral Fellowship. Dimitar Yonchev thanks the Translational Medicine Data Science Academy program at Novartis.

Author contributions

G.G. and R.R.P. conceived and supervised the study; S.D. led the data generation; D.Y. curated the dataset for modeling; M.B. implemented and evaluated the models; M.B., G.G. and R.R.P. discussed and analyzed the results; M.B. and R.R.P. wrote the manuscript; all authors revised the manuscript.

Competing Interests

The authors are/were employees and shareholders of Novartis AG.