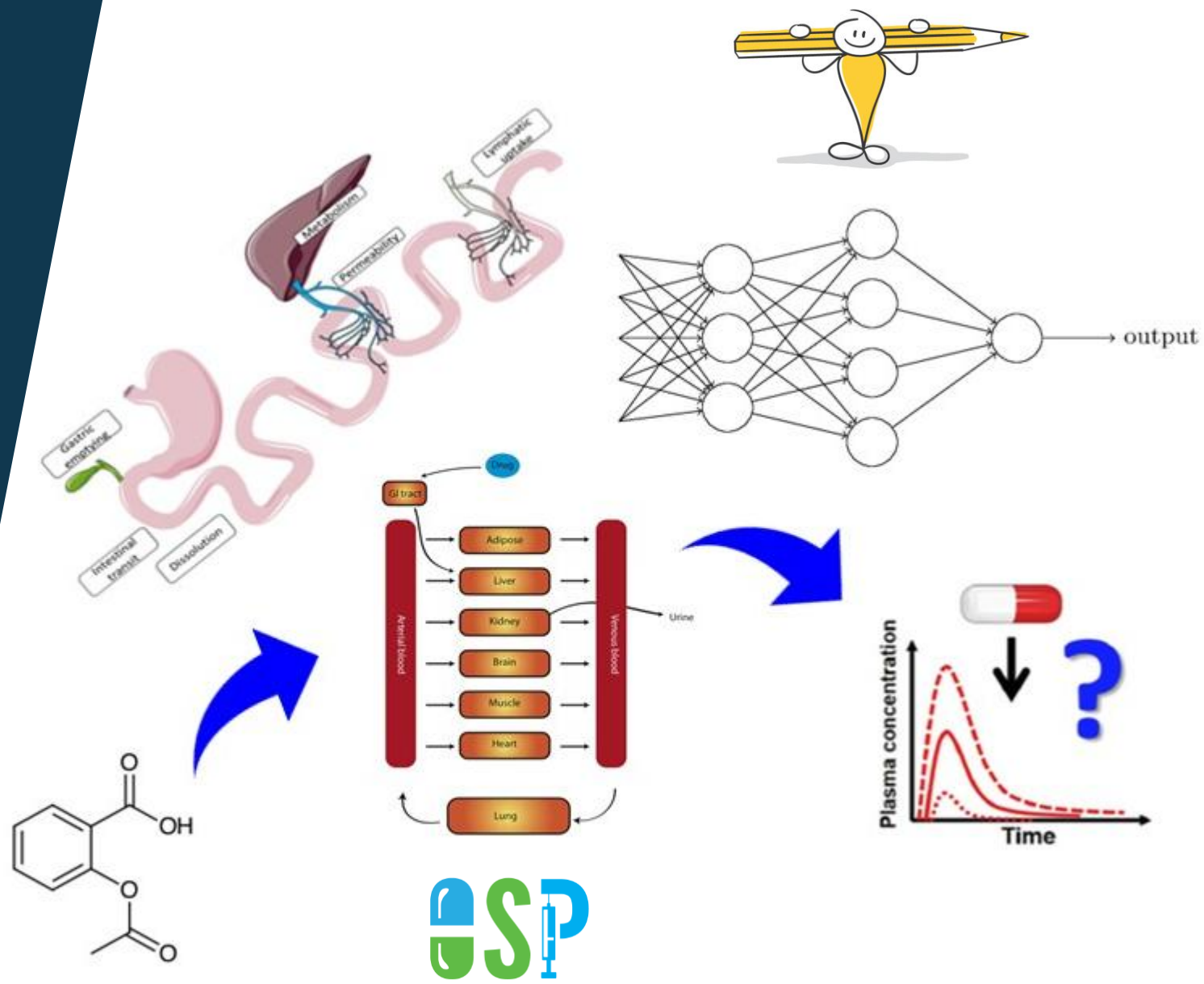
Insights on Predicting PK from Chemical Structure by Combining Machine Learning with Mechanistic Modeling

OSP conference 2025

2025/09/30

Andrea Gruber on behalf of the Bayer team
(Florian Führer, Stephan Menz, Holger Diedam,
Andreas H. Göller, Sebastian Schneckener)



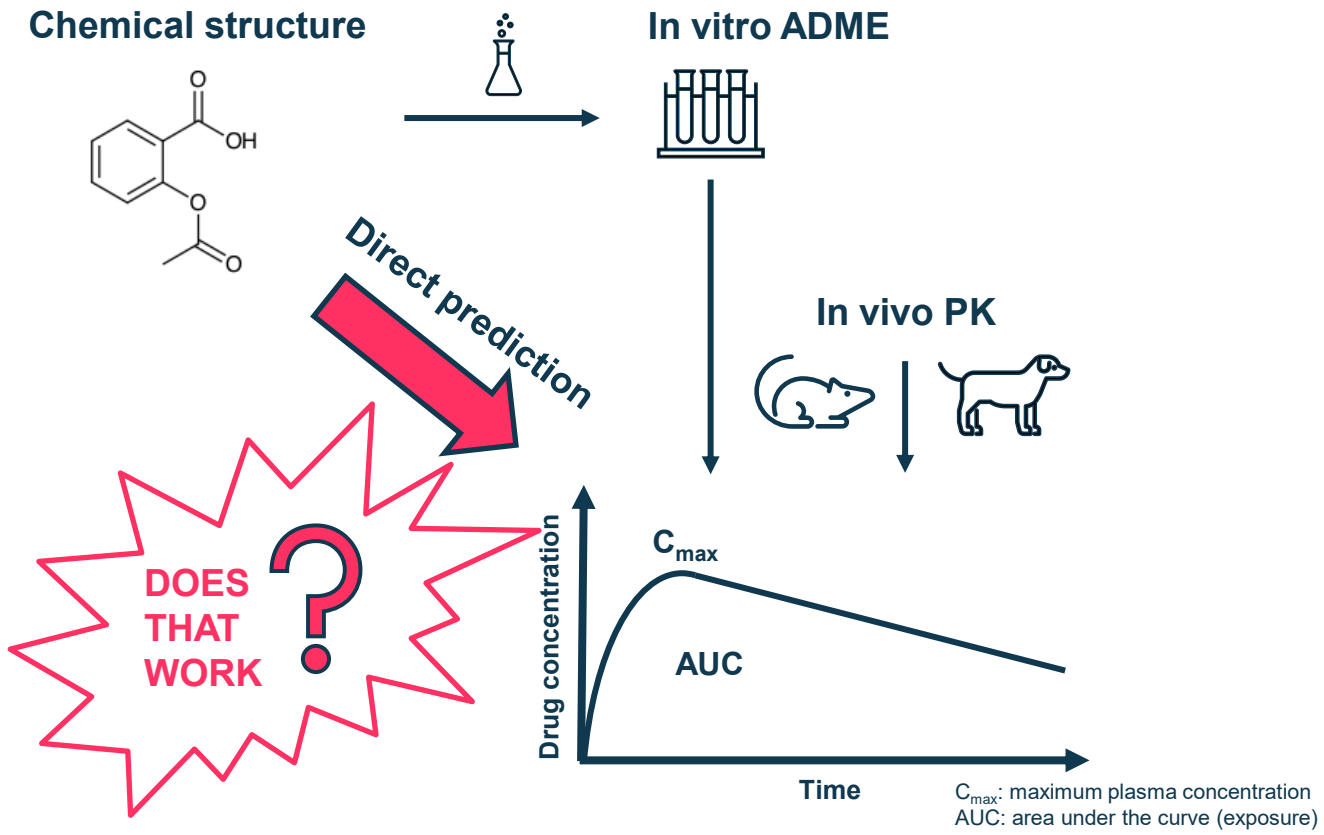
Understanding the Dose – Exposure – Response relationship

Pharmacokinetics (PK) | **Pharmacodynamics (PD)**

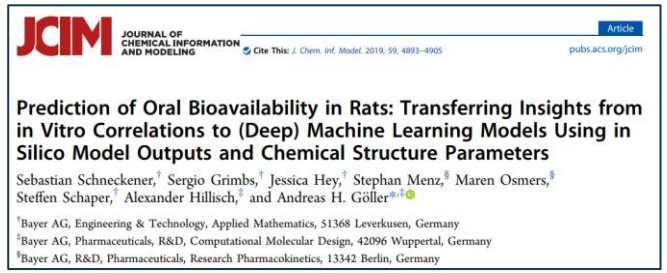
DOSE → ***Systemic Exposure*** (unbound plasma) ⇄ ***Target Exposure*** (unbound tissue) → ***Target Engagement*** → ***Target Modulation*** → ***Translation to Response***

In vitro ADME | **In vivo PK studies** | **In vitro PD** | **In vivo PD studies**

Motivation und Machine Learning model evolution



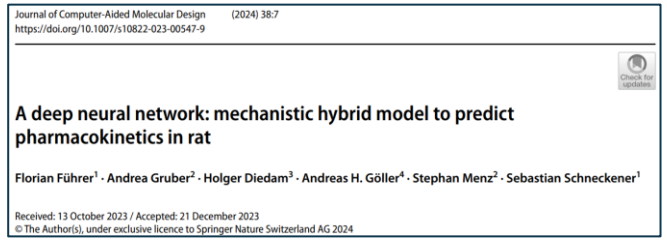
C_{max}: maximum plasma concentration
AUC: area under the curve (exposure)


JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

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

Sebastian Schneckener, Sergio Grimbs, Jessica Hey, Stephan Menz, Maren Osmers, Steffen Schaper, Alexander Hillisch, and Andreas H. Göller

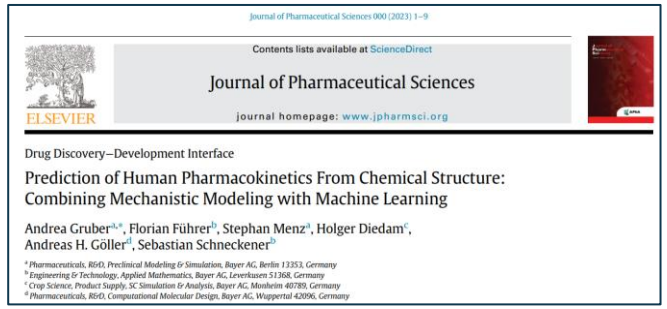
Schneckener et al. doi: 10.1021/acs.jcim.9b00460


Journal of Computer-Aided Molecular Design (2024) 38:7
https://doi.org/10.1007/s10822-023-00547-9

A deep neural network: mechanistic hybrid model to predict pharmacokinetics in rat

Florian Führer · Andrea Gruber · Holger Diedam · Andreas H. Göller · Stephan Menz · Sebastian Schneckener

Führer et al. doi: 10.1007/s10822-023-00547-9


Journal of Pharmaceutical Sciences

Drug Discovery–Development Interface

Prediction of Human Pharmacokinetics From Chemical Structure: Combining Mechanistic Modeling with Machine Learning

Andrea Gruber, Florian Führer, Stephan Menz, Holger Diedam, Andreas H. Göller, Sebastian Schneckener

Gruber et al. doi: 10.1016/j.xphs.2023.10.035


nature medicine

Explore content · About the journal · Publish with us

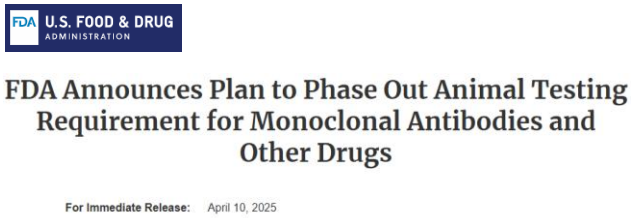
nature › nature medicine › news feature › article

News Feature | Published: 01 June 2023

Researchers and regulators plan for a future without lab animals

Sofia Moutinho

Nature Medicine 29, 2151–2154 (2023) | Cite this article


U.S. FOOD & DRUG ADMINISTRATION

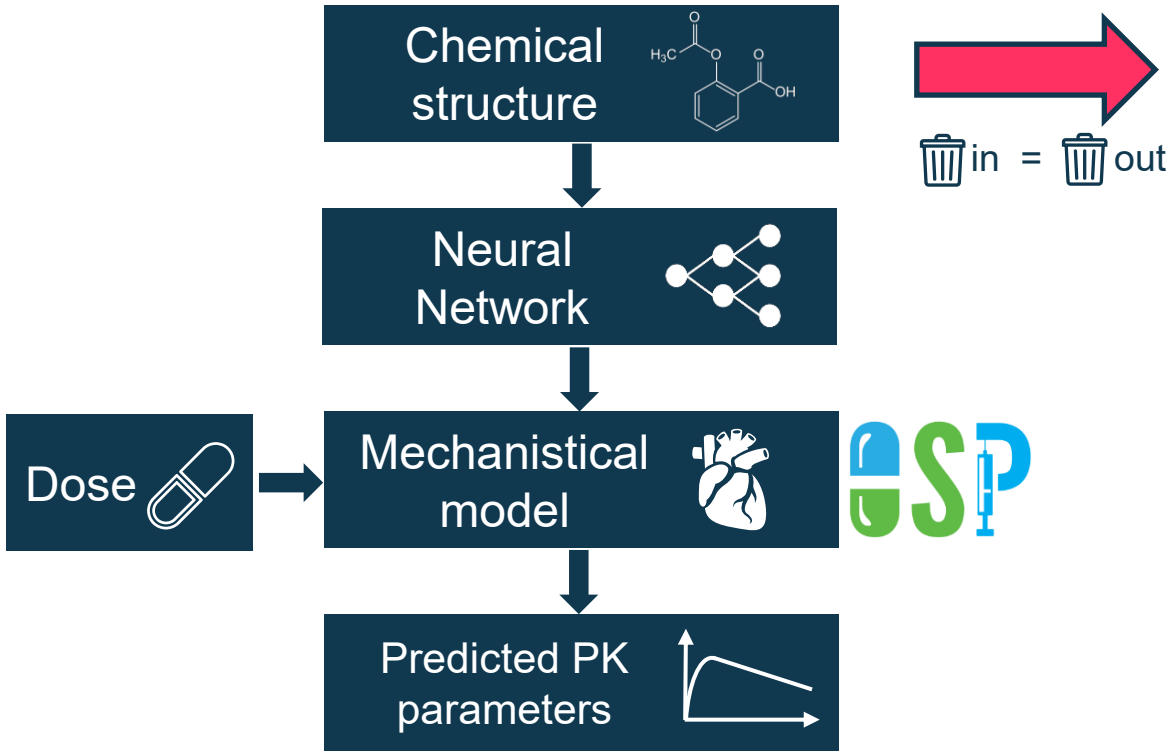
FDA Announces Plan to Phase Out Animal Testing Requirement for Monoclonal Antibodies and Other Drugs

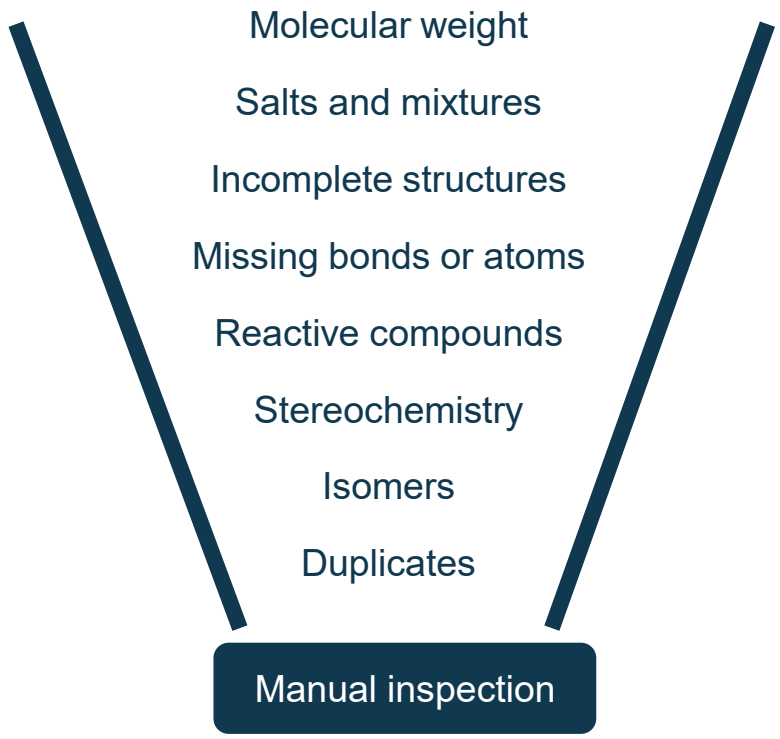
For Immediate Release: April 10, 2025

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

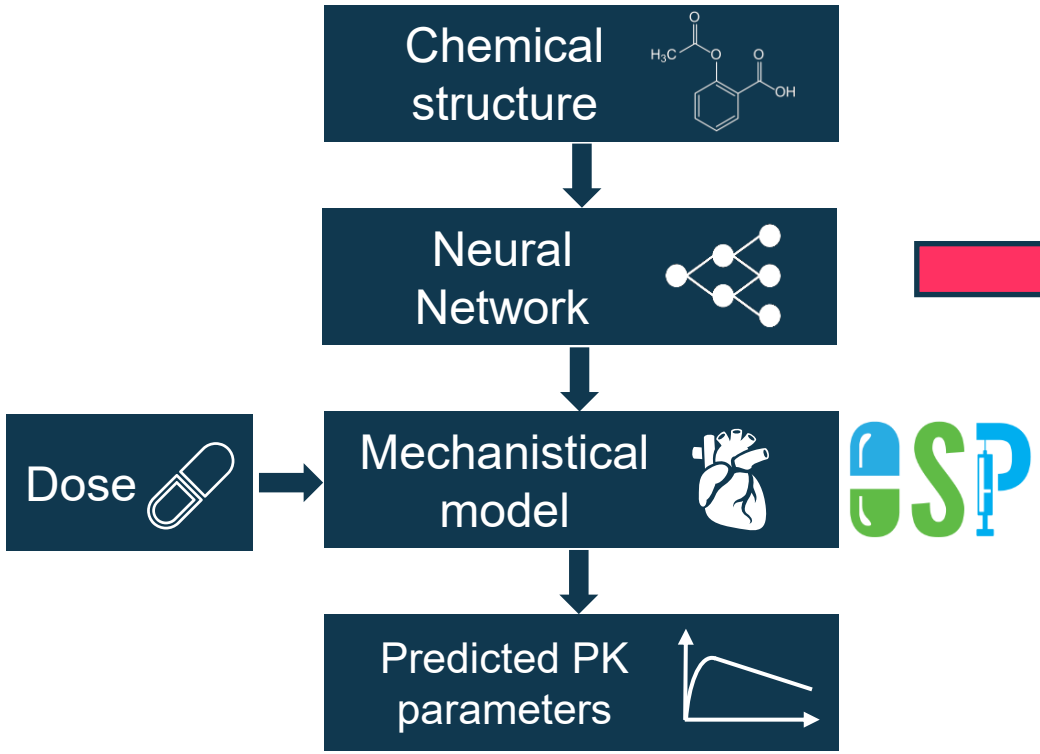
Regulatory acceptance of 3R (replacement, reduction, refinement) testing approaches - Scientific guideline

Hybrid model concept for rat and human PK prediction

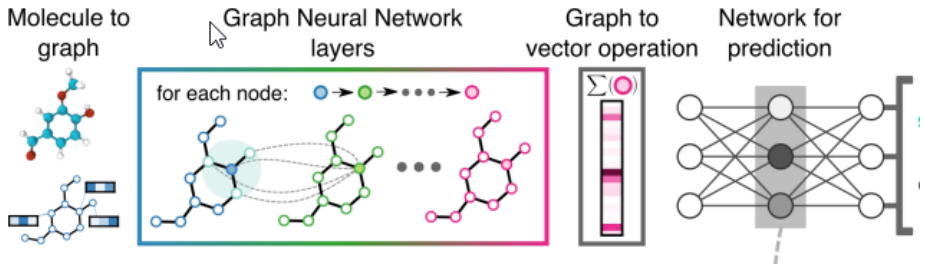




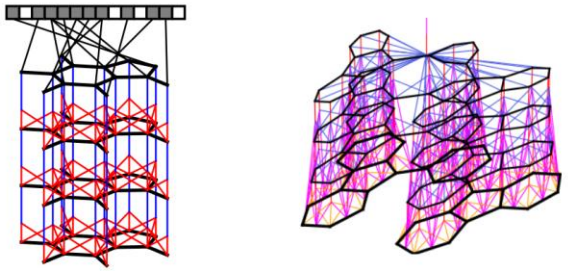
Hybrid model concept for rat and human PK prediction



Graph convolutional networks

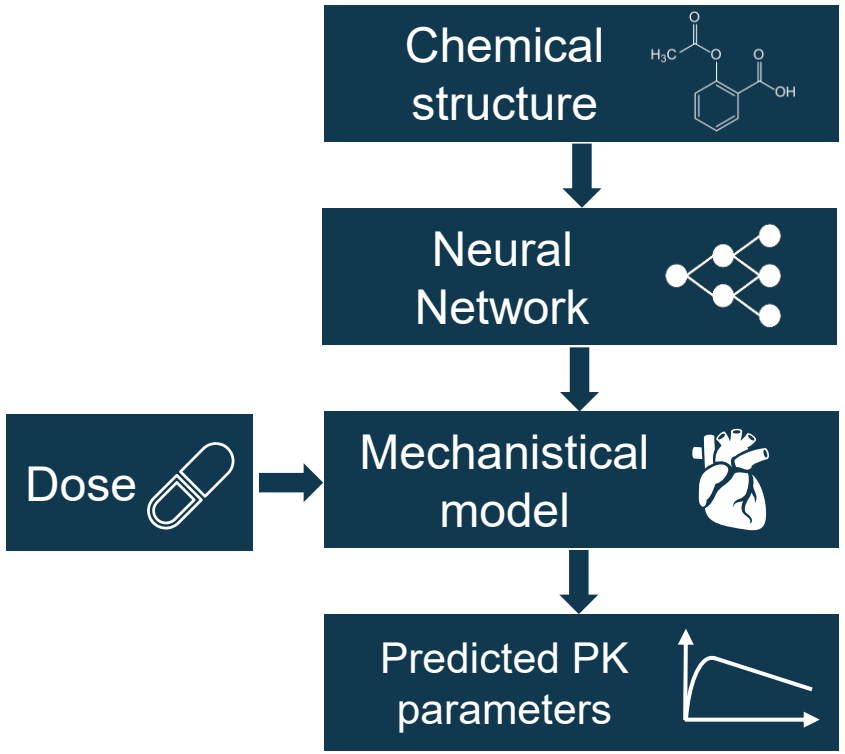


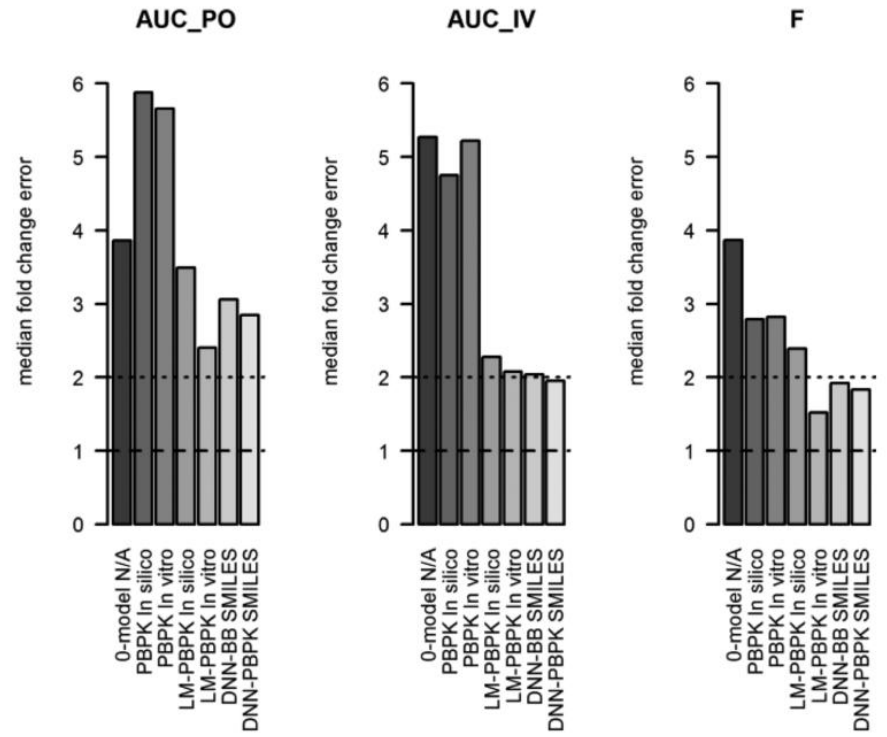
Sanchez-Lengeling B. et al. doi:10.48550/arXiv.1910.10685



Duvenaud D. et al. https://arxiv.org/pdf/1509.09292
Ramsundar B. et al https://books.google.de/books?id=tYFKuwEACAAJ.

Hybrid model concept for rat and human PK prediction

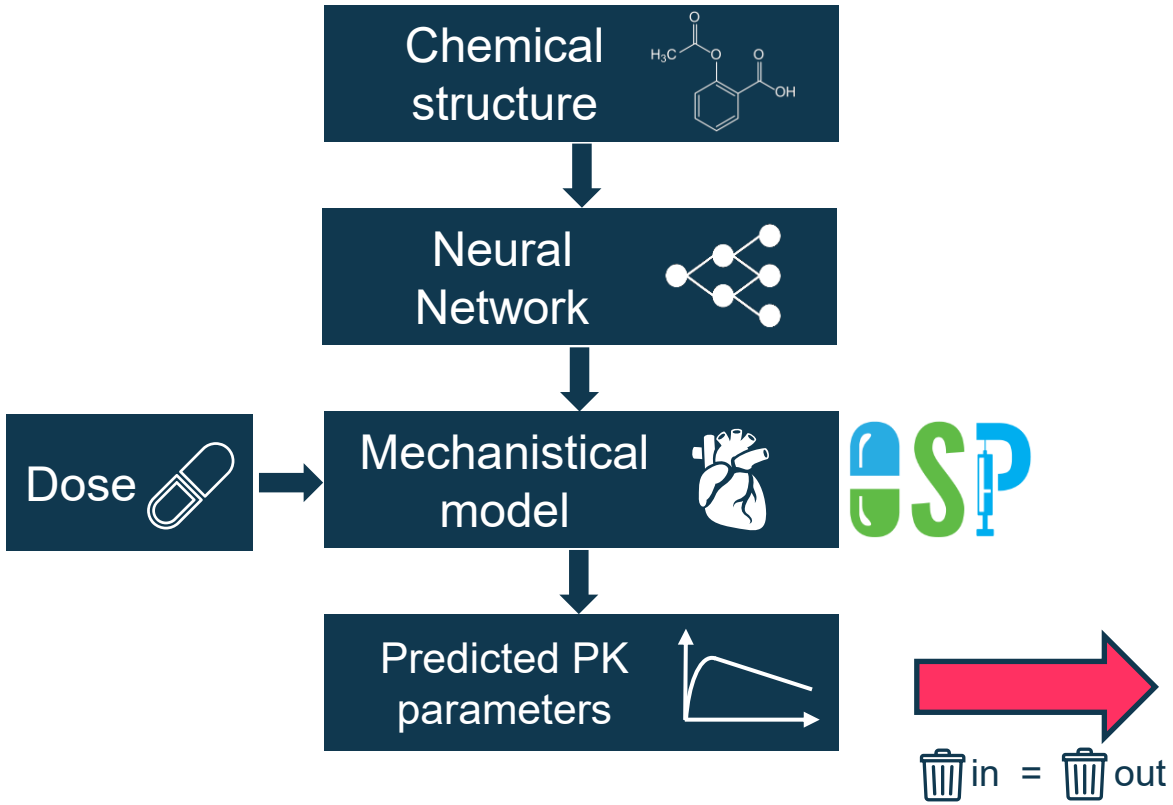




	model type						
	0-model	plain-PBPK	plain-PBPK	LM-PBPK	LM-PBPK	DNN-BB	DNN-PBPK
input type	N/A	in silico	in vitro	in silico	in vitro	SMILES	SMILES
AUC_{iv}	5.27	4.75	5.22	2.28	2.08	2.04	1.95
AUC_{po}	3.86	5.88	5.66	3.49	2.40	3.06	2.85
F	3.87	2.79	2.82	2.39	1.52	1.92	1.83

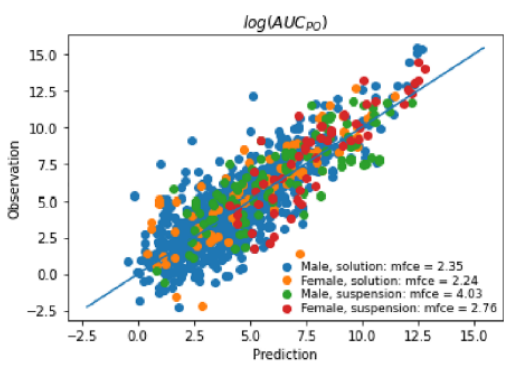
Schneckener et al. doi: 10.1021/acs.jcim.9b00460

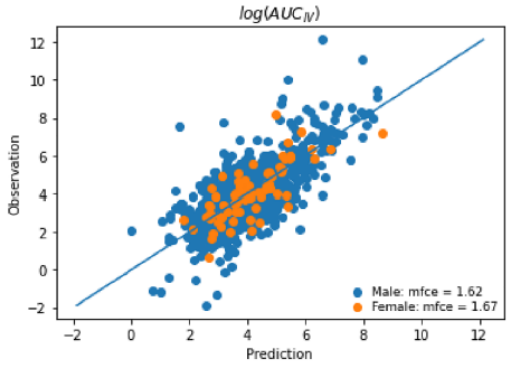
Hybrid model concept for rat and human PK prediction

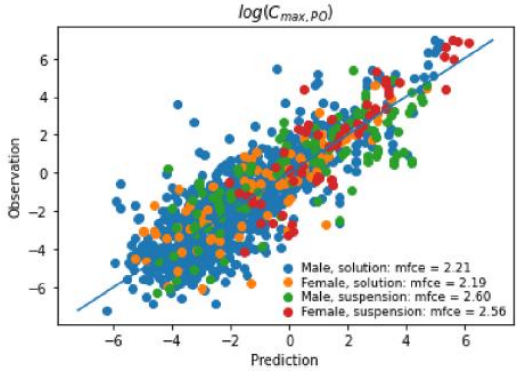


Data of ~7000 compounds from Bayer internal database	Data of ~3000 compounds from external databases Elsevier Reaxys and Cortellis Integrity
AUC_{iv}, AUC_{po}, $C_{max,po}$ data available in equal parts and across exposure classes (low, intermediate, high)	87% of datapoints from oral PK ($C_{max,po}$, AUC_{po}) vs 13% of intravenous PK (AUC_{iv}) with bias towards higher exposure data
Dose range up to 1000 mg/kg	Dose range up to 75 mg/kg with majority of data up to 10 mg/kg
Metadata available regarding sex and applied formulation	Metadata “health state” partly available in external databases

Rat hybrid model performance: evaluation on test data set



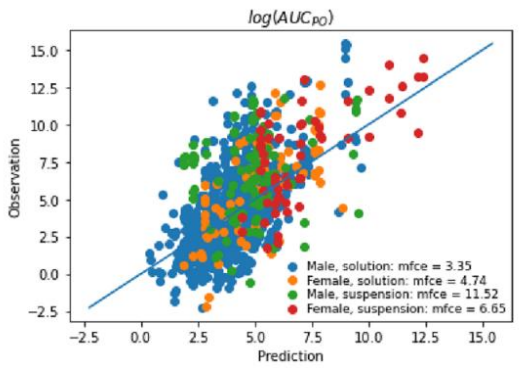


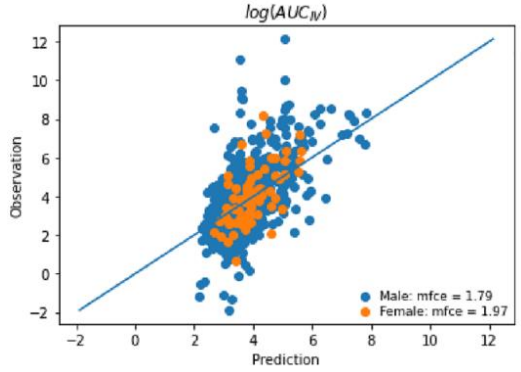


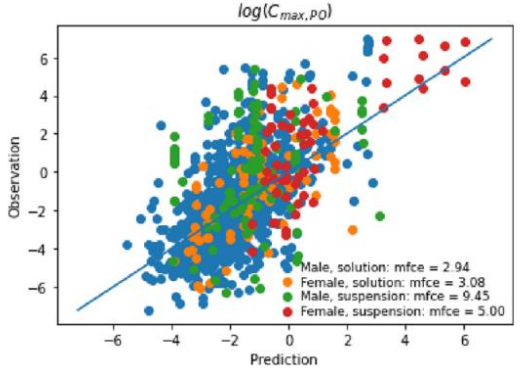
Median fold change error
mfce = exp(median |log(observation) – log(prediction)|)

- mfce = < 2 for AUC_{iv}
- mfce between 2.24 – 4.03 for AUC_{po}
- mfce between 2.19 – 2.6 for $C_{max,po}$

Improvement from previous SMILES-based Hybrid model



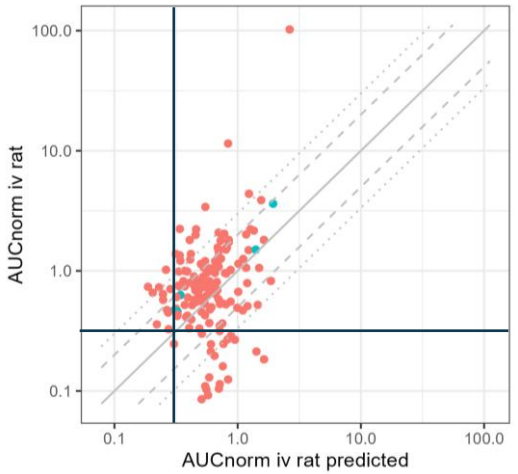




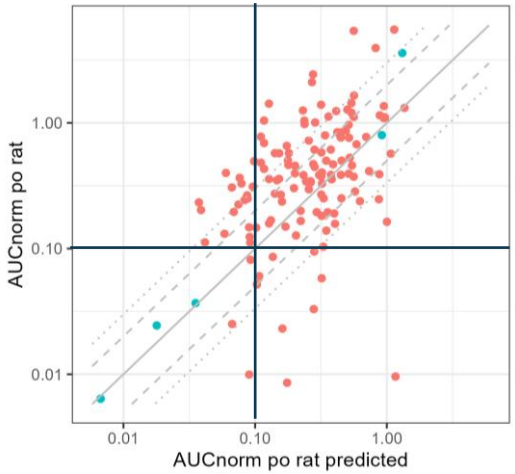
Comparison of hybrid model to pure deep learning model:
Higher accuracy of the hybrid model for all 3 endpoints

Rat hybrid model performance: evaluation on project level

Project example



Mfce = 1.76
Within 2-fold: 58%
Within 3-fold: 79%
Pearson correlation coefficient: 0.45
Spearman's rank correlation coefficient: 0.2



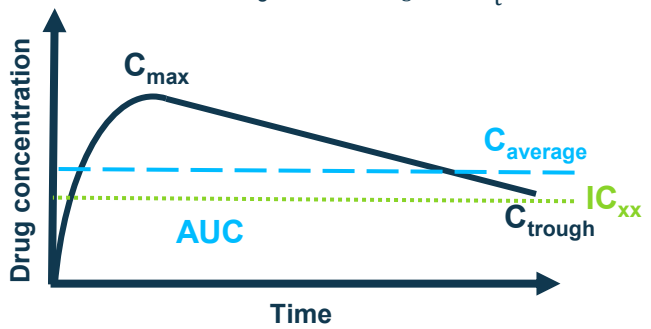
Mfce = 2.5
Within 2-fold: 50%
Within 3-fold: 70%
Pearson correlation coefficient: 0.47
Spearman's rank correlation coefficient: 0.51

- Compounds part of model training set
- New compounds
- Within 2-fold
- Within 3-fold

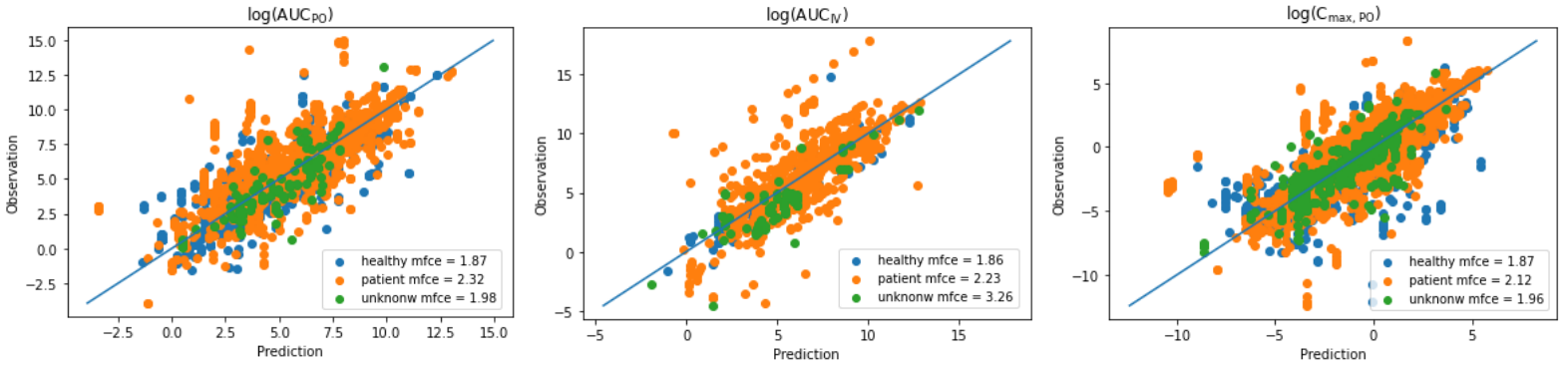
Project questions

Can we use the model for….

- prediction of clearance $\mathrm{CL} = \frac{1}{AUC_{norm\ iv}}$
- prediction of oral exposure AUC_{po}
- ranking of compounds regarding their predicted exposure
- classifying compounds into low / intermediate / high exposure
  → CL_{plasma} = CL_{blood} in relation to liver blood flow (<30%, 30-70%, >70%)
  - $AUC_{norm,iv}$: <0.34, 0.34-0.79, >0.79 kg*h/L
  - $AUC_{norm,po}$: <0.1, 0.1-0.55, >0.55 kg*h/L
- an early evaluation of developability (feasible dose)
  → Is the dose calculation based on efficacious AUC_{po} or C_{trough}/ICxx?
  - C_{trough} not directly predicted by the hybrid model
  - Full c-t profile simulation based on PBPK input parameters predicted from the hybrid model are possible but not directly accessible or mechanistically interpretable
  - Approximation of C_{trough} by $C_{average} = \frac{AUC_{po}}{\tau}$



Human hybrid model performance: evaluation on test data set



Median fold change error

mfce = exp(median |log(observation) – log(prediction)|)

- mfce between 1.86 – 3.26 for AUC_{iv}
- mfce between 1.87 – 2.32 for AUC_{po}
- mfce between 1.87 – 2.12 for $C_{max,po}$

	AUC_{po}	**AUC_{iv}**	**$C_{max,po}$**
Within 2-fold (%)	44	47	50
Within 3-fold (%)	62	65	68

- Predictions within 2- and 3-fold errors are comparable to published human PK prediction methods
  e.g.: Jones (2011), Davies (2020), Naga (2022), Fagerholm (2021), Miljković (2021)
- Direct comparison of hybrid model predictions for human PK to allometric scaling based on rat data showed similar predictive accuracy for AUC_{iv} (mfce = 2.48), but a strong benefit of the hybrid model for AUC_{po} predictions (mfce = 1.76 vs 2.9)
- Overall prediction accuracy for all exposure classes (low, intermediate, high) of 70 % (po) and 63 % (iv) with high AUCs showing precision of 73 % (iv) and 80 % (po)

Human hybrid model performance: simulation of c-t profiles

Open Systems Pharmacology · GitHub

OSP-PBPK-Model-Library Public

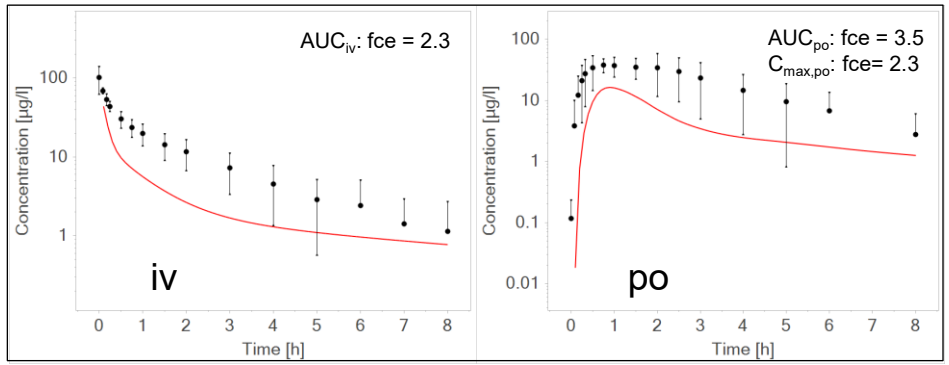
Library of released PBPK substance models and evaluation reports

MATLAB 31 24

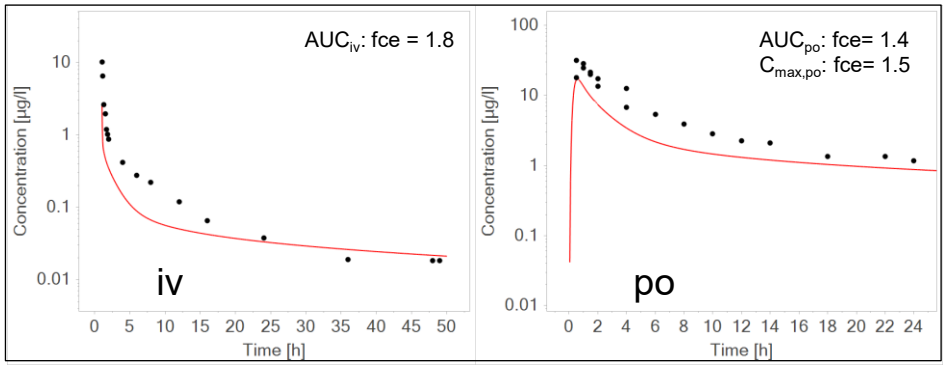
Compound examples from OSP library (human c-t profile data)

fce = |observation –prediction|

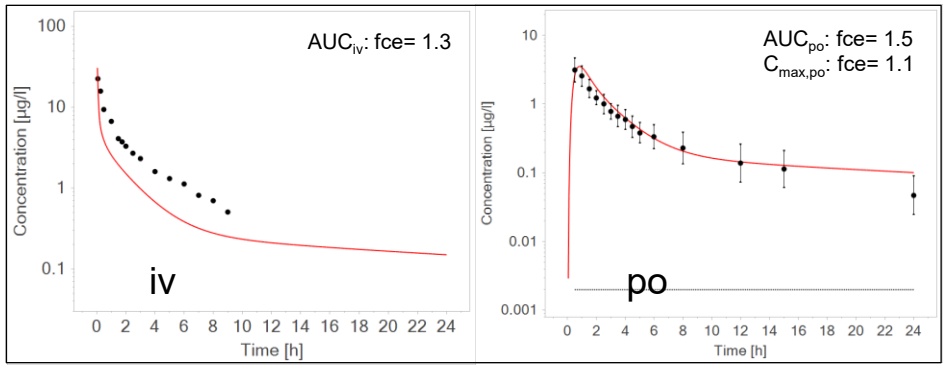
Alfentanil



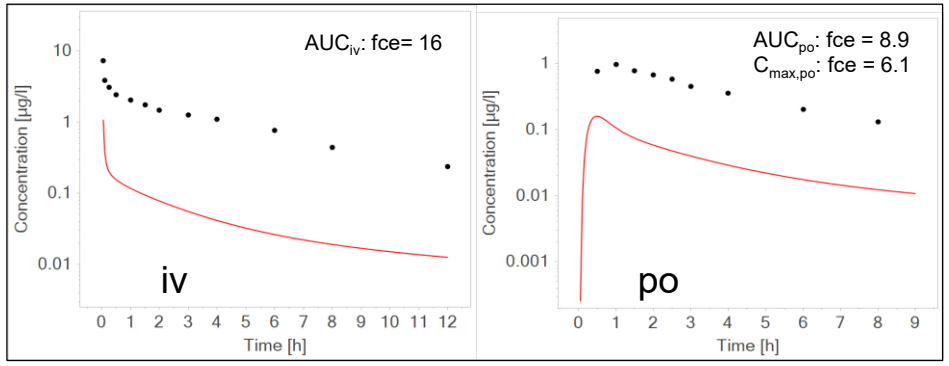
Dapagliflozin



Midazolam



Triazolam



Testing the extrapolation potential of the hybrid model to an endpoint it was not trained on → simulated c-t profiles in similar predictive accuracy as the trained endpoints, but resulting PBPK models not mechanistically meaningful

Conclusion & outlook

Further research currently ongoing for training on and predicting full c-t profiles in several preclinical species and increased chemical space

Models combine the mechanistic physiological knowledge of the PBPK models from the OSP suite with state-of-the-art Machine Learning for predictions within 2- to 3-fold accuracy

Rat and human hybrid model available for project work at Bayer

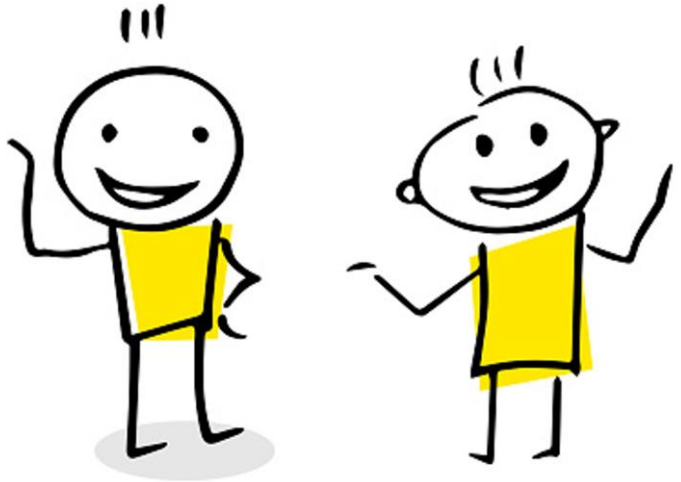
Rat model re-training proved to be very important for continuously high model performance and integration of new compound classes

Application of the hybrid models (dual screening) in early phases of discovery for filtering and prioritizing promising candidates for detailed PK characterization actively saves resources (time, expenses)

Consistent with the 3R principle: reduction, replacement and refinement of animal experiments

Thanks to the Team!

Bayer Pharma &
Bayer Crop Science



Hybrid modeling

Florian Führer
Stephan Menz
Holger Diedam
Andreas H. Göller
Sebastian Schneckener

Preclinical modeling and simulation

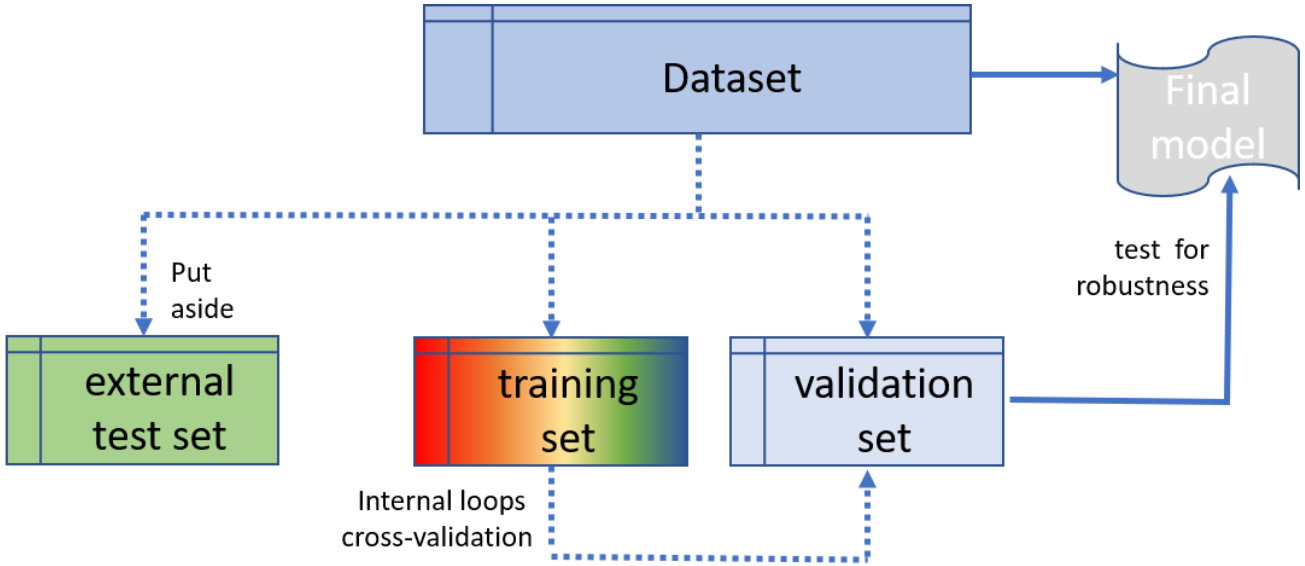
Andreas Reichel
Carsten Terjung
Christoph Hethey
Christoph Thiel
Darian Schirr
Filip Steinbauer
Jan-Erik Busse
Marcel Mischnik
Markus Krauss
Robin Haid

And many others who have contributed their time and resources to this cross-divisional initiative at Bayer

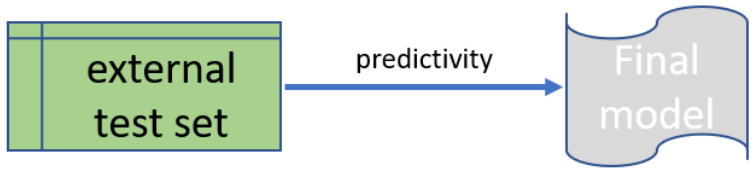
GMP – Good Modeling Practice

The real predicitivty of a model is assessed from a left out external data set

Model identification and internal validation



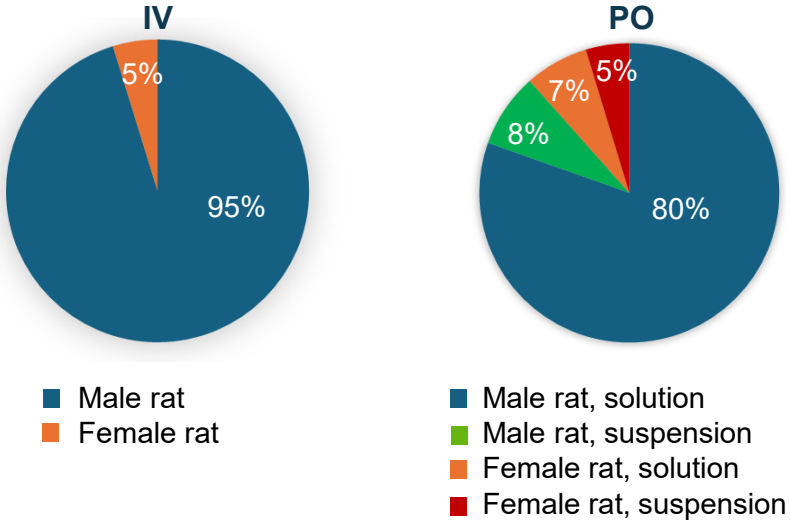
External validation



See Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. OECD Series on Testing and Assessment, No. 69, OECD Publishing: Paris, 2007.

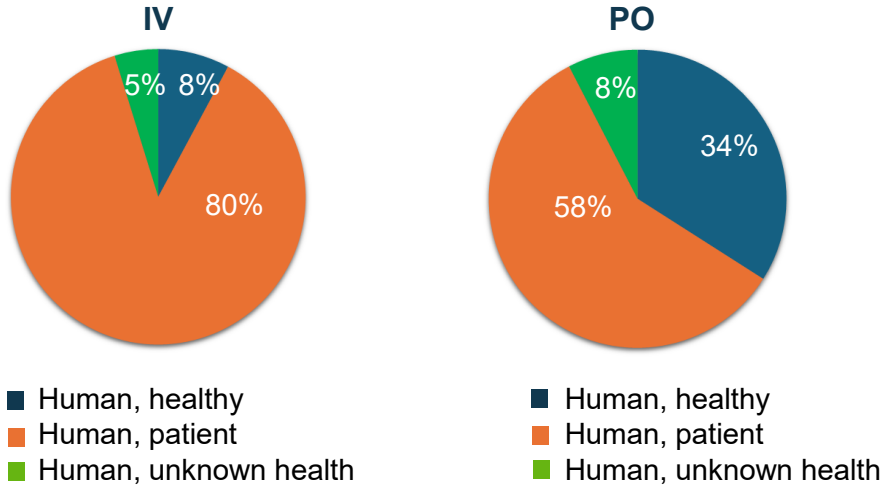
PK data for rat and human hybrid model

Data of **~7000 compounds**
from Bayer internal databases



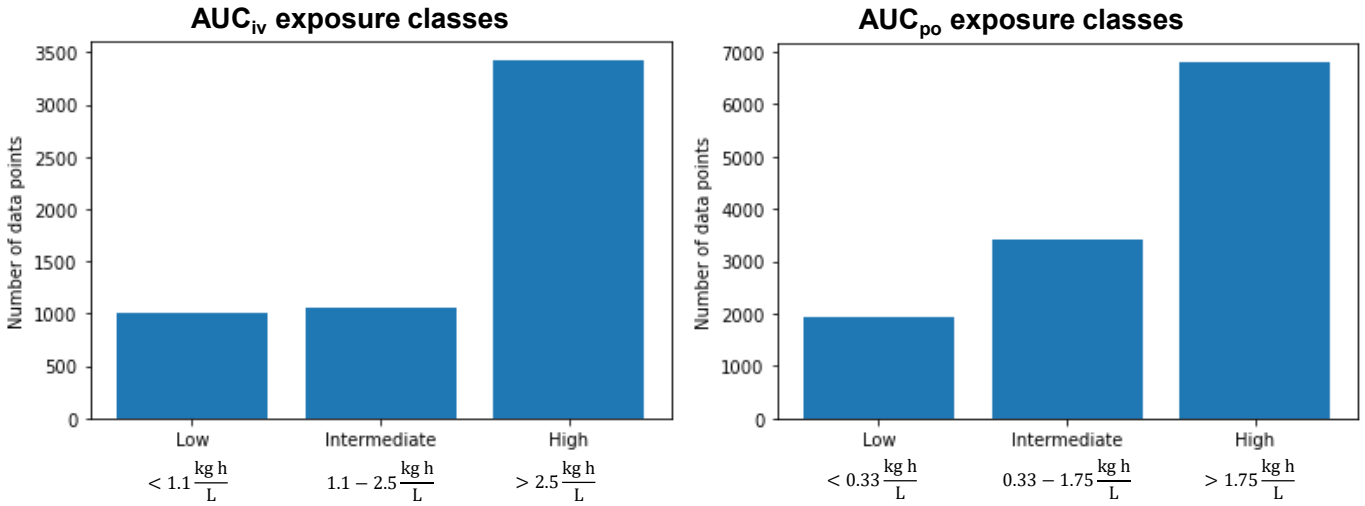
- AUC_{iv}, AUC_{po}, $C_{max,po}$ data available in equal parts
- Dose range up to 1000 mg/kg with PK and Tox studies in both low and high dose range (data from “Pharma” and “Crop Science” compounds)
- Metadata available regarding sex and applied formulation

Data of **~3000 compounds**
from external databases
Elsevier Reaxys and Cortellis Integrity

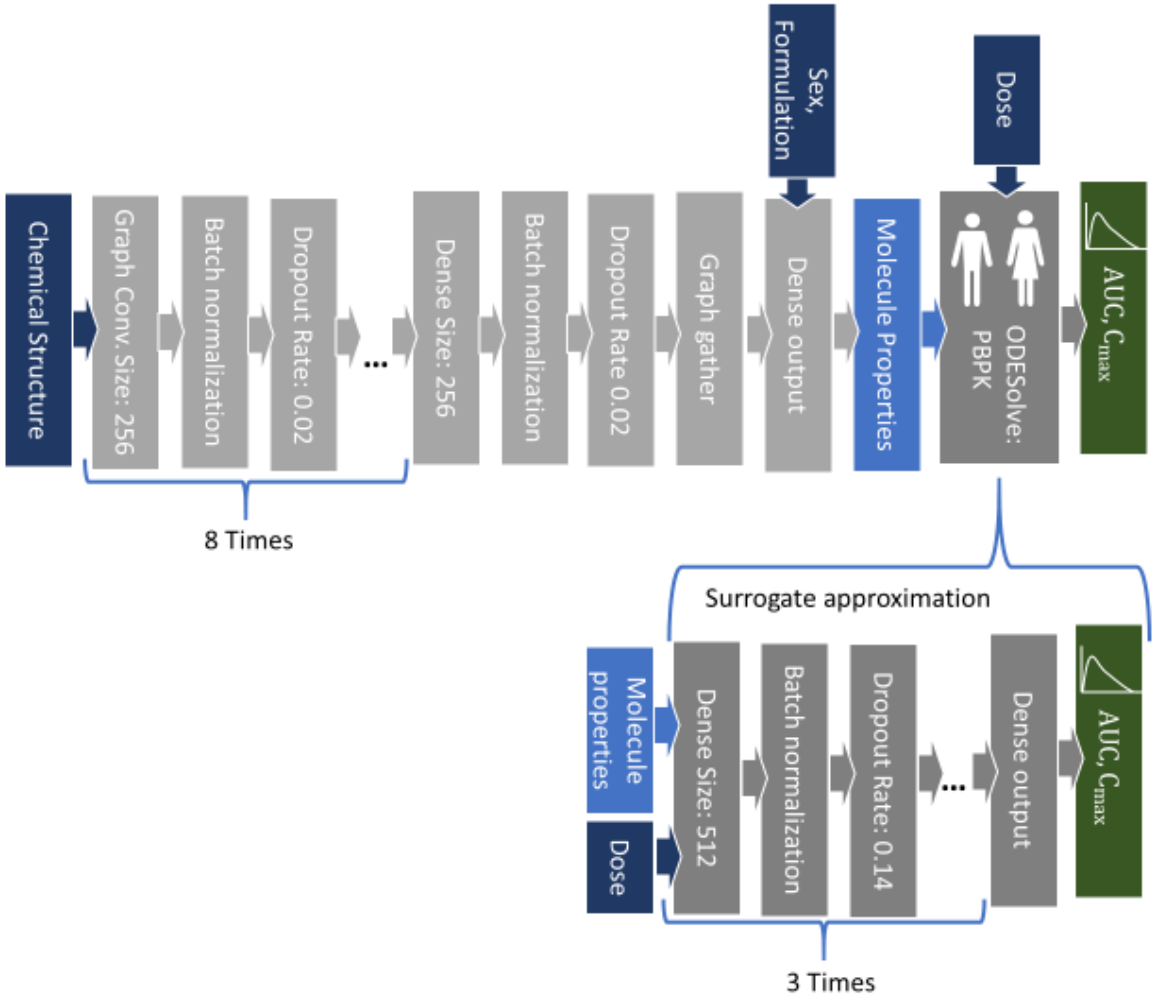


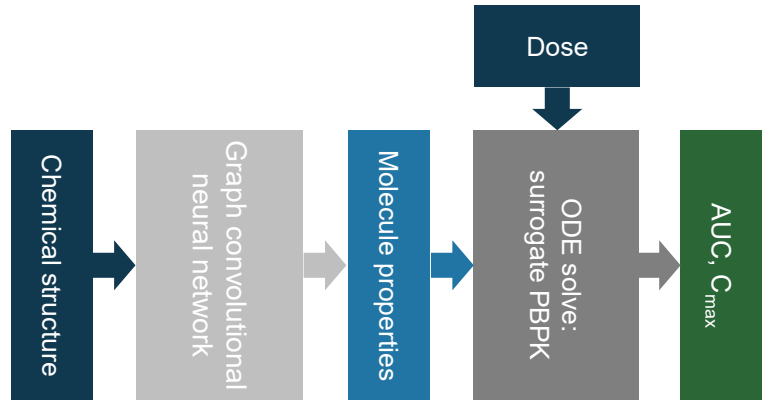
- ~87% of datapoints from oral PK ($C_{max,po}$, AUC_{po}) vs ~13% of intravenous PK (AUC_{iv})
- Dose range up to 75 mg/kg with majority of data up to 10 mg/kg
- Distinction between healthy subjects and patients not thoroughly possible on both databases

Human hybrid model input: bias towards higher exposure



Hybrid model structure details





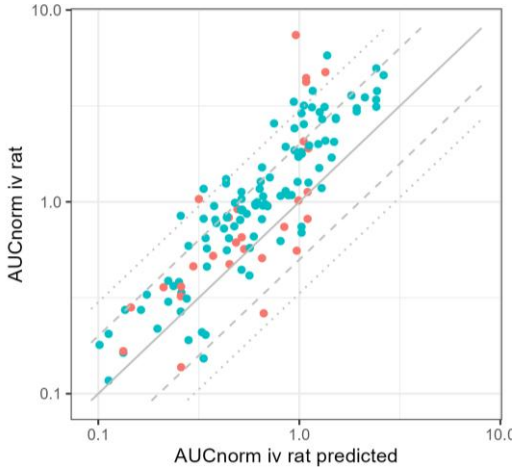
Mechanistic model parameter overview

Parameter	Data for pre-training *	Model/assay description
Hepatic clearance	*In vitro*	Hepatocyte stability assay
Vmax of P-gp–like active transport	*In vitro*	Caco-2 assay
Glomerular filtration rate (GFR)	No pre-training ** Random initialization	
Fraction unbound in plasma	*In silico*	Deep Learning model for humans
Lipophilicity	*In silico*	Deep Learning model for membrane affinity
Effective molecular weight	*In silico*	Molecular weight reduced by halogen contributions
Stomach solubility	*In silico*	Henderson-Hasselbach equation with reference solubility at pH=7 and pKa from Deep Learning models
Small intestine solubility	*In silico*	
Large intestine solubility	*In silico*	
Small intestine permeation	*In silico*	Predicted from membrane affinity and molecular weight
Large intestine permeation	*In silico*	

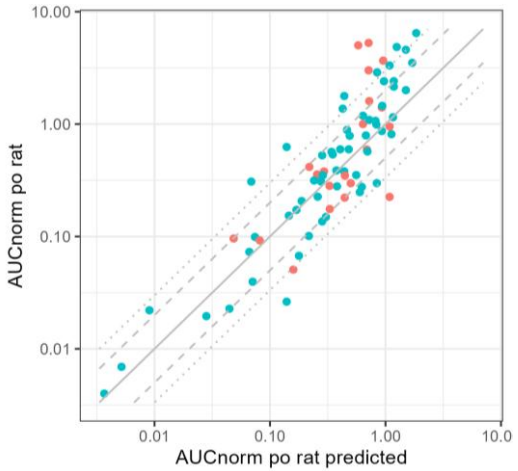
* Data used for pre-training is derived from Bayer internal *in vitro* assays and *in silico* models.
** Data for glomerular filtration rate (GFR) were not available, as determining the GFR would require urine data from *in vivo* trials. The corresponding output node of the property net is hence initialized randomly.

Rat hybrid model performance: additional project evaluations

Project example B

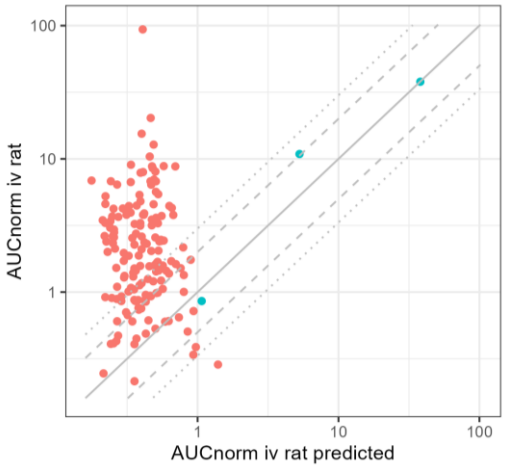


Mfce = 1.68
Within 2-fold: 74%
Within 3-fold: 90%
Pearson correlation coefficient:0.76
Spearman's rank correlation coefficient: 0.85

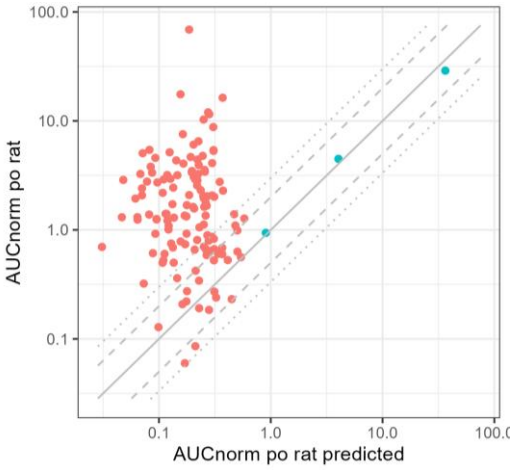


Mfce = 1.66
Within 2-fold: 63%
Within 3-fold: 80%
Pearson correlation coefficient:0.72
Spearman's rank correlation coefficient: 0.83

Project example C



Mfce = 8.99
Within 2-fold: 19%
Within 3-fold: 36%
Pearson correlation coefficient:0.33
Spearman's rank correlation coefficient: 0.0052

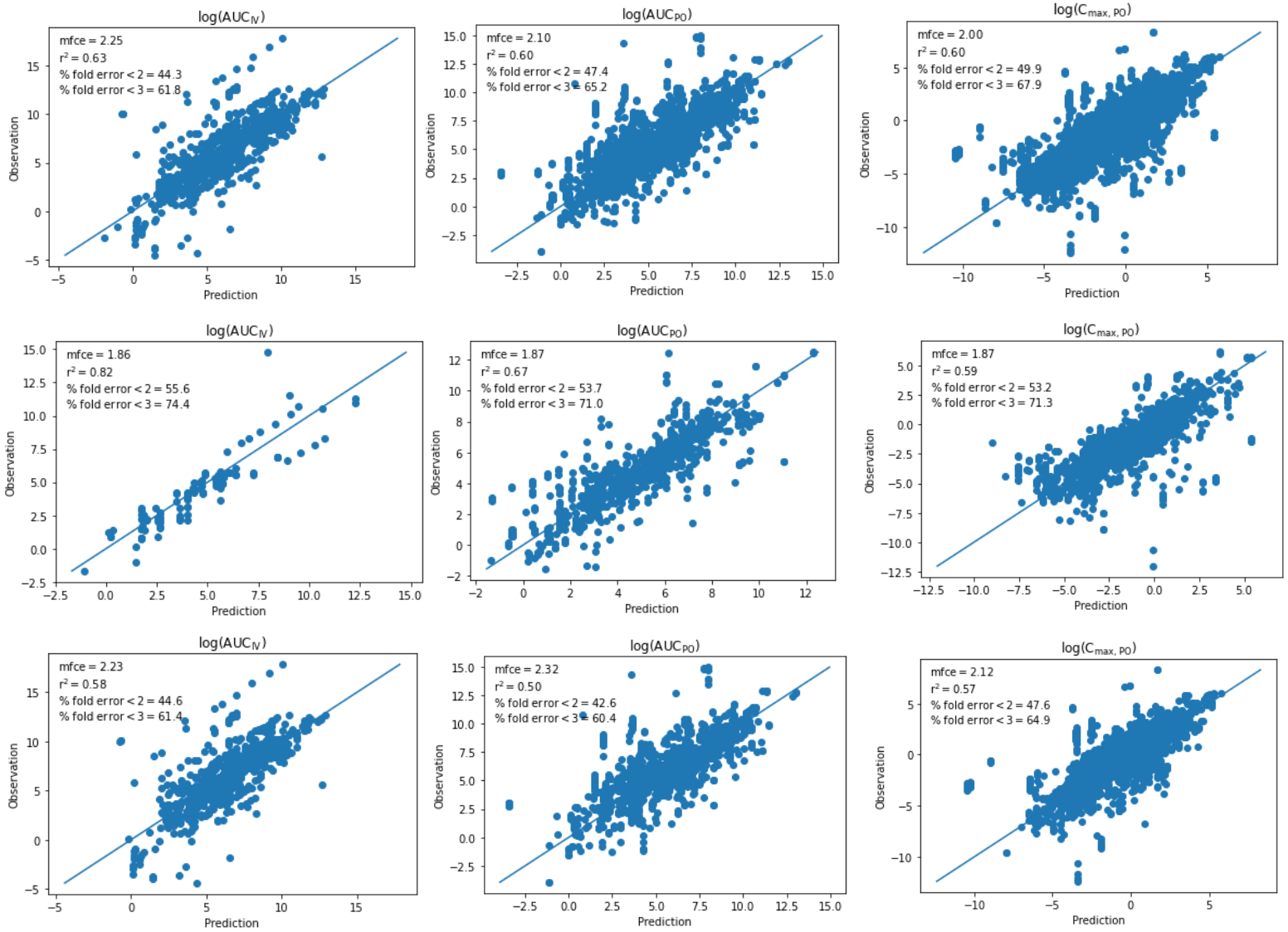


Mfce = 8.05
Within 2-fold: 17%
Within 3-fold: 28%
Pearson correlation coefficient:0.33
Spearman's rank correlation coefficient: 0.072

- Compounds part of model training set
- New compounds
- Within 2-fold
- Within 3-fold

Regular retraining of the hybrid model (1x / year) → new compound data can increase the prediction accuracy for ongoing projects, therefore directly impact project work and also increase the chemical space of the training data

Human hybrid model performance: evaluation on training data set



All subjects:

Mfce between 2.0 – 2.25

Healthy subjects

Mfce between 1.86 – 1.87

Patients

Mfce between 2.12 – 2.32

Comparison to published human PK prediction methods (selected examples)

Jones et al. doi: 10.2165/11539680-000000000-00000

Table IV. Pharmacokinetic prediction accuracy using physiologically based pharmacokinetic (PBPK) and compartmental approaches, using the predicted clearance (CL) as input

Prediction method	Prediction measure	V_{ss}, intravenous	CL, intravenous	AUC, oral	C_{max}, oral	t_{max}, oral	Terminal $t_{½}$, oral
GastroPlus™ PBPK	% within 2-fold error [3-fold error] of observed value	90 [100]	80 [85]	50 [72]	67 [72]	72 [94]	61 [83]
	Average fold error	1.4	1.6	2.7	2.0	1.7	2.1
1-compartment model	% within 2-fold error [3-fold error] of observed value	75 [85]	80 [85]	33 [56]	44 [61]	61 [78]	50 [61]
	Average fold error	1.6	1.6	3.9	2.5	1.9	2.5

AUC = area under the plasma concentration-time curve; **C_{max}** = maximum plasma concentration; **$t_{½}$** = half-life; **t_{max}** = time to reach C_{max}; **V_{ss}** = volume of distribution at steady state.

Miljković et al. doi: 10.1021/acs.molpharmaceut.1c00718

Table 2. Model Performance on a Hold-Out Test Set[a]

	N	R^2	RMSE	% <2-fold error	% <3-fold error	% <5-fold error
AUC PO	620	0.63	0.76	27.4	48.1	69.7
C_{max} PO	628	0.68	0.62	40.3	58.4	77.1
Vd_{ss} IV	103	0.47	0.50	48.5	68.0	77.7

[a]The performance statistics for AUC PO, C_{max} PO, and Vd_{ss} IV models on a hold-out test set are listed. For each model, number of tested compound−dose combinations, R^2, RMSE, and percentage of combinations within two-, three-, and fivefold error thresholds are reported.

Davies et al. doi: 10.1016/j.tips.2020.03.004

Table 2. Comparison of Percentages of AstraZeneca CDs with Predictions within Twofold of Observed Parameter Values (AUC, C_{max}, and $t_{1/2}$) with Other Reported Works

Evaluation	% CDs[b] with AUC predicted within twofold	% CDs[b] with C_{max} predicted within twofold	% CDs[b] with $t_{1/2}$ predicted within twofold
AZ 2000–2010	58% (46/79)	59% (34/58)	62% (42/68)
AZ 2011–2018	64% (18/28)	78% (18/23)	70% (19/27)
Van den Bergh *et al.* [67][a]	26–51%	46–63%	43–60%
Jones *et al* [68]	50% (9/18)	67% (12/18)	61% (11/18)
Zhang *et al* [66]	63% (10/16)	88% (14/16)	69% (11/16)

[a]Results given as ranges due to evaluation of a variety of methods (n = 35 CDs).
[b]Abbreviation: CD, candidate drug.

Naga et al. doi: 10.1021/acs.molpharmaceut.2c00040

Table 1. Error Metrics of the IV Parameters Predictions for the Six Different Simulations

parameter	error metric	(1) direct scaling	(2) dilution	(3) unbound	(4) back-calculated	(5) machine learning[a]	(6) Austin
CL (mL/min/kg) (n = 432)	% 2fe	57.6	41.7	22.5	98.8	35.9	33.3
	% 3fe	76.4	63	38.9	100	60.2	50.9
	AFE	1.42	0.463	0.212	1	0.476	0.302
	AAFE	2.05	2.53	4.81	1.13	2.76	3.48
	RMSLE	0.842	1.02	1.46	0.165	1.1	1.24
	CCC(log)	0.398	0.423	0.309	0.981	0.176	0.397
	ρ	0.471	0.541	0.528	0.98	0.246	0.574
	$R2$	0.179	0.198	0.181	0.952	0.0391	0.217
	$R2$(log)	0.222	0.33	0.379	0.964	0.0902	0.419
AUC_{inf} (ng· h/mL) (n = 432)	%2fe	57.6	41.4	22.9	98.8	36.1	33.3
	%3fe	76.4	63	38.9	100	60.2	50.9
	AFE	0.703	2.16	4.71	1	2.1	3.31
	AAFE	2.05	2.53	4.81	1.14	2.76	3.48
	RMSLE	0.949	1.15	1.86	0.187	1.22	1.53
	CCC(Log)	0.603	0.545	0.364	0.986	0.422	0.464
	ρ	0.6222	0.638	0.564	0.982	0.489	0.611
	$R2$	0.0782	0.216	0.401	0.974	0.129	0.353
	$R2$(log)	0.419	0.471	0.436	0.972	0.308	0.489
V_{ss} (L/kg) (n = 423)	% 2fe	59.1	60	60.8	59.8	45.4	60.5
	% 3fe	81.6	82	82.3	81.3	70.4	82
	AFE	0.692	0.702	0.704	0.694	1.01	0.703
	AAFE	2.01	2	2	2.02	2.45	2
	RMSLE	0.538	0.538	0.539	0.542	0.663	0.539
	CCC(Log)	0.582	0.584	0.584	0.576	0.412	0.584
	ρ	0.603	0.602	0.602	0.598	0.46	0.602
	$R2$	0.449	0.447	0.446	0.425	0.29	0.447
	$R2$(log)	0.401	0.4	0.399	0.392	0.182	0.399

[a]Machine learning column also uses ML for f_{up} and Log D not just for clearance.

Table 3. Error Metrics of the Oral Parameter Prediction for the Six Different Simulations

parameter	error metric	(1) direct scaling (n = 479)	(2) dilution (n = 480)	(3) Austin (n = 480)	(4) back-calculated CL + in vitro physchem (n = 480)	(5) ML physchem + back-calculated CL (n = 480)	(6) ML (all properties) (n = 480)
AUC_{inf} (ng ·h/mL)	% 2fe	38	31.9	23.3	59.4	63.5	27.9
	% 3fe	56.8	50.4	40.8	80	81.9	45.4
	AFE	0.589	2.62	4.13	0.79	0.905	2.9
	AAFE	3.29	3.57	4.8	2.12	2.01	4.2
	RMSLE	1.53	1.6	1.93	1.1	1.03	1.8
	CCC(Log)	0.559	0.55	0.502	0.801	0.825	0.417
	ρ	0.6	0.673	0.662	0.855	0.858	0.512
	$R2$	0.075	0.254	0.229	0.384	0.497	0.475
	$R2$(log)	0.367	0.473	0.477	0.654	0.682	0.322
C_{max} (ng/mL)	% 2fe	40.5	38.8	36.9	47.5	48.1	33.5
	% 3fe	58	59	54.6	72.5	66.2	50.4
	AFE	0.884	2.13	2.51	1.03	1.53	2.41
	AAFE	2.97	3.12	3.34	2.46	2.53	3.69
	RMSLE	1.36	1.45	1.54	1.16	1.21	1.65
	CCC(Log)	0.563	0.549	0.532	0.713	0.715	0.453
	ρ	0.561	0.618	0.622	0.755	0.758	0.531
	$R2$	0.111	0.206	0.273	0.359	0.447	0.133
	$R2$(log)	0.32	0.395	0.408	0.514	0.555	0.289
F_{oral}	% 2fe	66.3	68.6	68.6	64.5	68.1	65.9
	% 3fe	84.9	85.4	85.2	83	84.7	82.7
	AFE	0.83	1.22	1.26	0.808	0.928	1.46
	AAFE	1.89	1.85	1.88	2.05	1.95	1.94
	RMSLE	0.844	0.824	0.836	0.959	0.909	0.873
	CCC(lin)	0.0607	0.0515	0.0491	0.0724	0.0743	0.0205
	ρ	0.307	0.257	0.221	0.309	0.307	0.157
	$R2$	0.0227	0.0161	0.0142	0.0241	0.0253	0.00425
	$R2$(log)	0.0477	0.0218	0.018	0.0547	0.053	0.0016

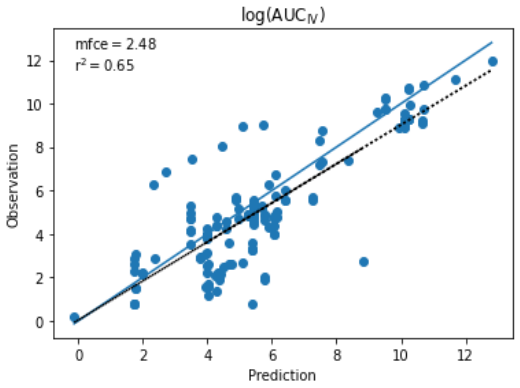
Human hybrid model performance: comparison to allometric scaling (rat)

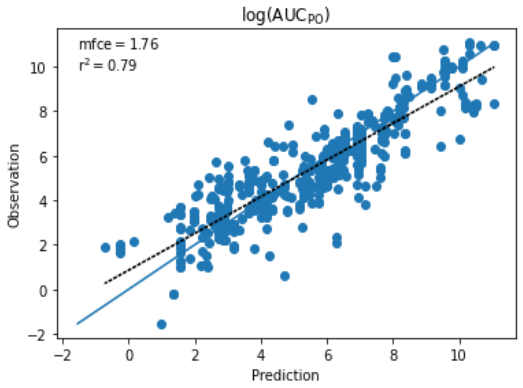
Selected test set with both rat and human PK data

AUC_{iv} prediction

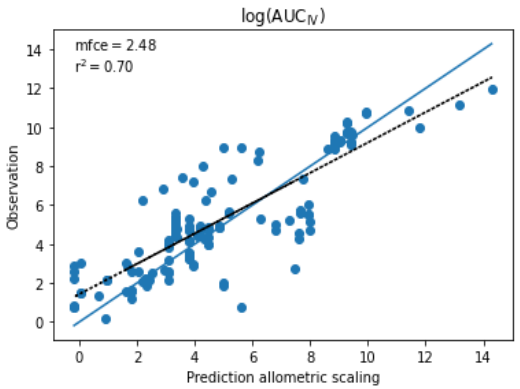
AUC_{po} prediction

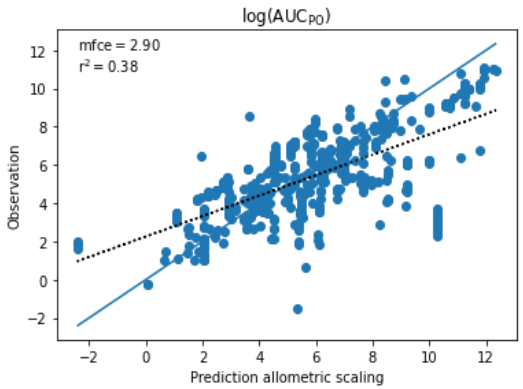
Hybrid model





Allometric scaling





Model comparison

Hybrid model vs allometric scaling:

- Allometric scaling based on single species scaling from rata data performed on selected test set with both rat and human data

$$\mathrm{CL_{human}} = \mathrm{CLanimal} * \left(\frac{\mathrm{BW_{human}}}{\mathrm{BW_{animal}}}\right)^{b}$$

BW_{human} = 73 kg, BW_{animal} = 0.23 kg, allometric scaling exponent b = 0.75

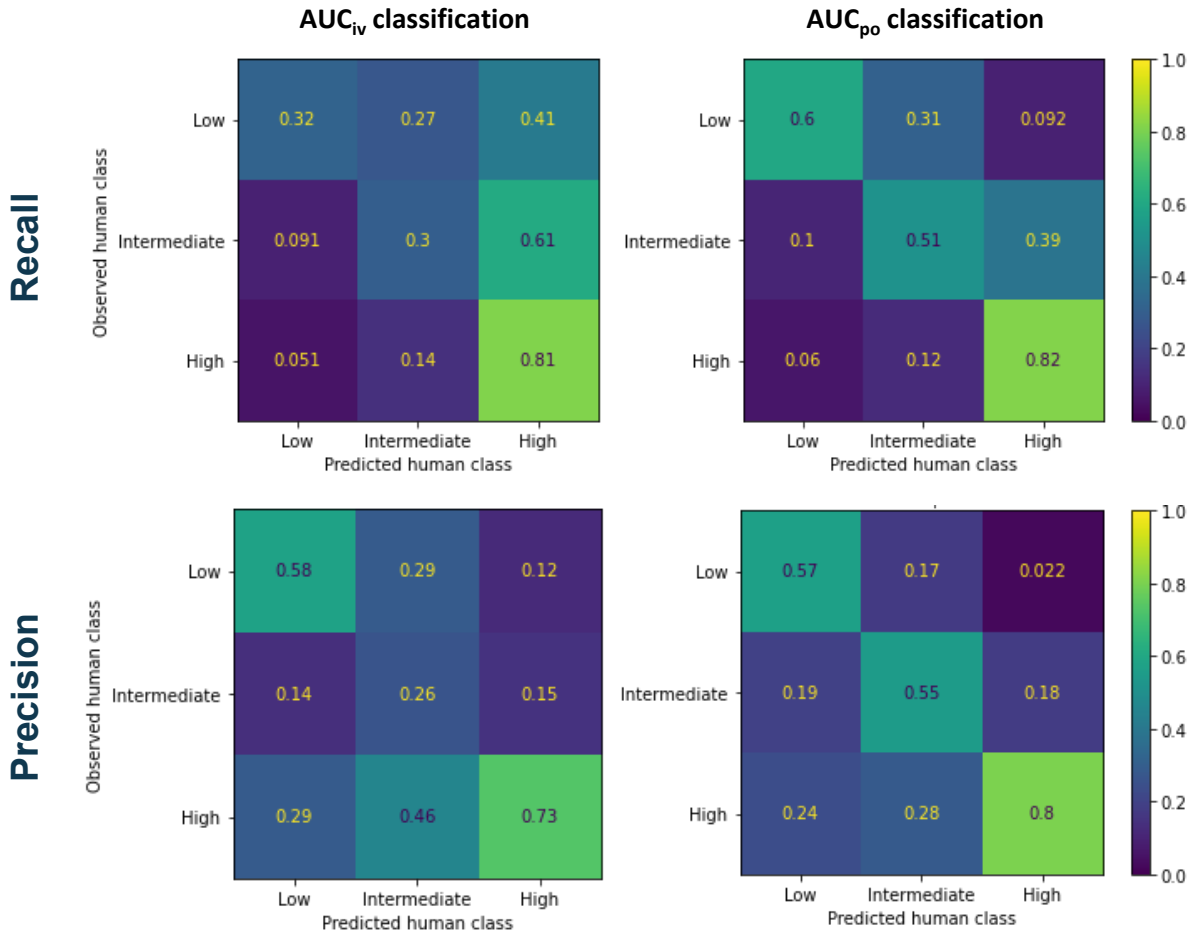
- AUC_{iv} is predicted by both methods with an mfce = 2.48
- AUC_{po} is predicted better by the hybrid model vs allometric scaling: mfce = 1.76 vs 2.9

→ Allometric scaling is a valid and standard method to predict human clearance and volume of distribution, but assumptions for bioavailability and oral absorption strongly impact the human PK prediction after oral dosing

→ The hybrid model has learned to account for these processes more efficiently and can deliver better predictions for AUC_{po}

Human hybrid model performance: evaluation of exposure classes

Confusion matrices showing model sensitivity (recall) and model precision



Confusion matrices

Recall
Recall, also known as **Sensitivity** and **True Positive Rate**, answers the question: "Of all the actual positive cases, how many did the model correctly identify?".

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

Precision
Precision is a metric that answers the question: "Of all the positive predictions made by the model, how many were actually correct?". It is a ratio of true positive predictions out of all positive predictions made by the model.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

Model performance analysis on AUC_{iv} data in structure-based clusters

- Clustering was performed on 5493 data points of AUC_{iv} data in Pipeline Pilot 2023 using ECFP-4 fingerprints (50 clusters)
- Number of data points per cluster < 500
- Training data set shows very balanced learning for all clusters (~around 1 log unit)
- More similar distributions and prediction performances in clusters of the test set vs training set for the larger clusters (e.g., 7, 28 or 31)
- Larger differences and worse prediction performances in the test set vs training set in clusters containing fewer compounds (e.g., cluster 40 or 21)