

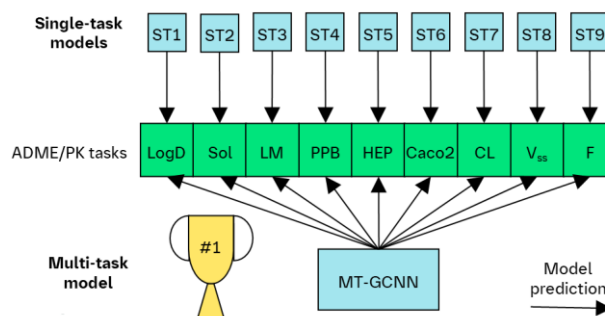
Multi-task ADME/PK Prediction at Industrial Scale: Leveraging Large and Diverse Experimental Datasets

Moritz Walter^a, Jens M. Borghardt^b, Lina Humbeck^{a*}, Miha Skalic^{a*}

[a] Medicinal Chemistry Department, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88397 Biberach an der Riss, Germany.

[b] Drug Discovery Sciences Department, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, 88397 Biberach an der Riss, Germany.

ABSTRACT: ADME (Absorption, Distribution, Metabolism, Excretion) properties are key parameters to judge whether a drug candidate exhibits a desired pharmacokinetic (PK) profile. In this study, we tested multi-task machine learning (ML) models to predict ADME and animal PK endpoints trained on in-house data generated at Boehringer Ingelheim. Models were evaluated both at the design stage of a compound (i.e., no experimental data of test compounds available) and at testing stage when a particular assay would be conducted (i.e., experimental data of earlier conducted assays may be available). Using realistic time-splits, we found a clear benefit in performance of multi-task graph-based neural network models over single-task models, which was even stronger when experimental data of earlier assays is available. In an attempt to explain the success of multi-task models, we found that especially endpoints with the largest numbers of data points (physicochemical endpoints, clearance in microsomes) are responsible for increased predictivity in more complex ADME and PK endpoints. In summary, our study provides insight into how data for multiple ADME/PK endpoints in a pharmaceutical company can be best leveraged to optimize predictivity of ML models.



INTRODUCTION

A successful drug needs to combine a range of desirable properties. Of major relevance for both efficacy and safety of a compound is its pharmacokinetic (PK) characteristics. The term PK comprises how the drug is absorbed, distributed, metabolized, and finally excreted from the body, i.e., the ADME properties. To estimate whether a drug might possess a desirable PK profile in human, several experiments are typically conducted in a drug discovery project. These range from simple physico-chemical measurements (e.g., logD) over *in vitro* ADME assays (e.g., plasma protein binding) to *in vivo* animal PK studies. Since some of these experiments are either expensive and time-consuming, or more importantly require animal experiments and/or (tissue) samples from animals / humans, *in silico* predictions for ADME-endpoints coming from machine learning (ML) models gained a lot of attention in recent years.^{1,2} These models learn from existing experimental data by linking chemical features to activities and are referred to as QSAR or QSPR models (quantitative structure-activity/property relationship).³ A wide range of different chemical descriptors and ML algorithms have been tested for QSAR modeling,⁴⁻⁶ aiming to predict physicochemical properties and *in vitro* ADME,⁷⁻¹⁵ PK in animals,^{7,9,16,17} or PK in humans.^{18,19}

Traditionally, compounds have been represented with chemical features such as substructural fingerprints²⁰ or calculated molecular properties²¹ and typical learning algorithms like support vector machine (SVM)²², random forest (RF)²³ and gradient tree boosting (e.g., XGBoost)²⁴ were used to train QSAR models. In recent years, deep neural network architectures gained more popularity.²⁵ These architectures enable different featurizations of molecules, for instance learned representations from chemical graphs (i.e., atoms as nodes, bonds as edges) and have been implemented in different modeling toolkits.²⁶⁻²⁸ Among those, Chemprop has been widely used in recent years for ADME and PK prediction^{10,11,16,19,29,30}. Learned chemical representations as those from Chemprop may outperform models based on classical chemical descriptors, yet this depends on the modelled dataset.³¹ Moreover, neural networks can be trained on several related endpoints at the same time resulting in so-called multi-task models.³² Earlier studies used traditional chemical descriptors in combination with multi-task feedforward neural networks and they reported a benefit of multi-task models over single task ones.^{12,14} In the study by Wenzel et al., R^2 (coefficient of determination) scores for prediction of microsomal clearance (in eight

species) on a test set were all improved in comparison to single task models. On the other hand, model performance for predicting LogD and Caco2 permeability was worse for multi-task models. Two studies from Novartis reported a successful use of multi-task Chemprop models. The prediction of brain penetration could be improved (R^2 0.52 vs 0.39) when trained together with several auxiliary tasks for *in vitro* permeability and lipophilicity²⁹. In the second study, microsomal stability in several species was predicted with multi-task Chemprop showing benefits for all endpoints in comparison to the tested single task model (XGBoost and single-task Chemprop).¹⁰ The focus of a different study was the prediction of *in vivo* rat PK parameters (AUC₀₋₂₄, F, CL, C_{max}, t_{1/2} and V_{ss}).³⁰ They tested several algorithms (including single-task and multi-task Chemprop) and used *in vitro* ADME properties (predicted or measured) as input to the models. In this scenario, the multi-task Chemprop models trained on all *in vivo* parameters overall showed no clear benefit compared to single-task Chemprop models trained on the parameters individually, demonstrating that a benefit of multi-task modeling is not guaranteed in all situations. More studies are required to fully understand the benefits and limitations of multi-task QSAR models.

In the present study, we investigated multi-task modeling on in-house *in vitro* ADME and *in vivo* PK datasets in comparison to single task models in a realistic setting with a prospective validation scheme. We also tested whether experimental data of early conducted assays (e.g., rat microsomal stability) can boost the performance when predicting related, but more complex endpoints like *in vivo* clearance in rats. Some details on the order how experiments are typically conducted (i.e., screening cascades), are provided below. Moreover, we analyzed which auxiliary assays are most useful in a multi-task model to improve the performance for other tasks. Our study demonstrates how multi-task modeling can be successfully used for ADME and PK prediction in drug discovery.

MATERIAL AND METHODS

Study overview

For this study, we used data for 28 endpoints (physicochemical properties, *in vitro* ADME and *in vivo* PK) from Boehringer Ingelheim's internal compound database. An overview of included endpoints is provided in Table 1. The endpoints were assigned to four sequential tiers (Tier 0 to Tier 3) to reflect the order, in which the assays typically are carried out in drug discovery projects in so-called screening cascades. For these screening cascades, some best practice exists, e.g., that *in vitro* experiments are performed before *in vivo* experiments to ensure tolerability of administered doses in PK studies. Another example is that *in vitro* ADME assays are executed first e.g., for the first PK or PD species and human, and only if deemed meaningful are profiled for other preclinical species.

The assays in Tier 0 (logD at pH 2 and 11) are part of the purification process of compounds. Assays in Tier 1 are cost-effective and straightforward tests that enable the evaluation of a large quantity of compounds in a high-throughput manner. These include aqueous solubility, as well as metabolic stability in liver microsomes. Endpoints of Tier 2 include more complex *in vitro* assays: metabolic stability in hepatocytes; binding to plasma protein; permeability and efflux in the Caco2 cell line. Tier 3 contains *in vivo* PK evaluations in rat or mice from single compound and/or cocktail studies with up to 5 compounds: total clearance (i.v. dosage); volume of distribution at steady state (i.v. dosage); and bioavailability. Note that this a generic description of a typical screening cascade, and in practice individual projects may deviate from it (e.g., not all assays from one tier are necessarily conducted at the same time).

Endpoints in different tiers may relate to the same physiological processes, with different level of biological complexity. For example, microsomal stability only represents metabolism by specific metabolizing enzymes (such as hepatic CYP enzymes), whereas hepatocytes also include transporter-mediated uptake into the hepatocytes as well as phase-I and phase-II metabolism and additional soluble enzymes (e.g., aldehyde oxidase). In contrast, the *in vivo* clearance provides a holistic evaluation by also comprising potential extra-hepatic metabolism as well as renal excretion. *In vivo* endpoints, such as bioavailability can even be considered an interplay of multiple *in vitro* ADME characteristics, i.e., the permeability and solubility impact the absorbed

fraction from the gastro-intestinal tract, and the hepatic metabolism impacts the hepatic first pass. Therefore, the assumption is that *in silico* predictions for later stage *in vitro* or *in vivo* experiments will improve when including prior *in vitro* endpoints.

The goal of this study was to train ML models that make best use of all available data. It was tested how multi-task models trained on several related endpoints at the same time compare to conventional single task models in a realistic setting in a pharmaceutical company. We employed temporal splitting (based on registration date, i.e., the date the compound is added to the internal database) to enable a realistic evaluation of the models. For multi-task models we further distinguish between performance at the design stage and at the testing stage of compounds. Figure 1 illustrates how the different scenarios relate to a typical workflow in a drug discovery project. The DMTA cycle is an established framework in drug discovery which states that new compounds are designed by chemists, synthesized, tested, and analyzed with insights then motivating new designs. With evaluation at design stage, we mean the compound is still virtual and no experimental data is available for it. Hence, the respective ML model can only be trained on data for previously synthesized compounds. As mentioned above, a screening cascade defines in which order different experiments are conducted for a compound. In practice, this means that experimental data of earlier tiers may be available when experiments of later tiers are conducted. Multi-task models may incorporate this information to make better predictions (for details see the description of the ML methods). Further details on the splitting strategy to reflect the described scenarios are given in the section "Model evaluation" below.

Table 1 Overview of included endpoints.

Tier	Endpoint	Data points (train-val-test)	unit
0	LogD at pH=2 (LogD2)	119k-28k-39k	-
	LogD at pH=11 (LogD11)	126k-26k-34k	-
1	High-throughput solubility at pH=2.2 (HTSol2.2)	86k-5k-4k	μmol/L
	High-throughput solubility at pH=4.5 (HTSol4.5)	112k-7k-8k	μmol/L
	High-throughput solubility at pH=6.8 (HTSol6.8)	112k-7k-7k	μmol/L
	Metabolic stability in human liver microsome assay (hLM)	125k-8k-9k	%QH ^a
	Metabolic stability in rat liver microsome assay (rLM)	65k-5k-7k	%QH ^a
	Metabolic stability in mouse liver microsome assay (mLM)	49k-6k-8k	%QH ^a
2	Binding to human plasma protein (hPPB)	13k-2k-3k	%bound
	Binding to rat plasma protein (rPPB)	10k-2k-1k	%bound
	Binding to mouse plasma protein (mPPB)	5k-1k-1k	% bound
	Metabolic stability in human hepatocyte assay with 0% serum (hHEP0)	1k-<1k-<100	%QH ^a
	Metabolic stability in human hepatocyte assay with 5% serum (hHEP5)	4k-<1k-2k	%QH ^a
	Metabolic stability in human hepatocyte assay with 50% serum (hHEP50)	5k-<1k-<1k	%QH ^a
	Metabolic stability in rat hepatocyte assay with 0% serum (rHEP0)	<1k-<100-<100	%QH ^a
	Metabolic stability in rat hepatocyte assay with 5% serum (rHEP5)	9k-1k-<1k	%QH ^a
	Metabolic stability in rat hepatocyte assay with 50% serum (rHEP50)	5k-<1k-<1k	%QH ^a
	Metabolic stability in mouse hepatocyte assay with 0% serum (mHEP0)	<1k-<100-<100	%QH ^a
	Metabolic stability in mouse hepatocyte assay with 5% serum (mHEP5)	4k-<1k-1k	%QH ^a
	Metabolic stability in mouse hepatocyte assay with 50% serum (mHEP50)	5k-<1k-<1k	%QH ^a
	Permeability in Caco2 cell line (Caco Perm)	17k-4k-4k	10-6 cm/sec
	Efflux ratio in Caco2 cell line (Caco Efflux)	15k-3k-4k	Efflux ratio
	3	<i>In vivo</i> clearance in rat (rCL)	10k-<1k-<1k
<i>In vivo</i> clearance in mouse (mCL)		4k-<1k-<1k	mL/(min*kg)
Volume of distribution at steady state in rat (rVss)		10k-<1k-<1k	L/kg
Volume of distribution at steady state in mouse (mVss)		5k-<1k-<1k	L/kg
Oral bioavailability in rat (rF)		3k-<1k-<100	%
Oral bioavailability in mouse (mF)		2k-<1k-<1k	%

^a %QH describes the percentage of liver blood flow cleared. For different species the following liver blood flows are assumed: human 20.7 mL/min/kg, rat: 70 mL/min/kg, mouse: 90 mL/min/kg.

The goal of this study was to train ML models that make best use of all available data. It was tested how multi-task models trained on several related endpoints at the same time compare to conventional single task models in a realistic setting in a pharmaceutical company. We employed temporal splitting (based on registration date, i.e., the date the compound is added to the internal

database) to enable a realistic evaluation of the models. For multi-task models we further distinguish between performance at the design stage and at the testing stage of compounds. Figure 1 illustrates how the different scenarios relate to a typical workflow in a drug discovery project. The DMTA cycle is an established framework in drug discovery which states that new compounds are designed by chemists, synthesized, tested, and analyzed with insights then motivating new designs. With evaluation at design stage, we mean the compound is still virtual and no experimental data is available for it. Hence, the respective ML model can only be trained on data for previously synthesized compounds. As mentioned above, a screening cascade defines in which order different experiments are conducted for a compound. In practice, this means that experimental data of earlier tiers may be available when experiments of later tiers are conducted. Multi-task models may incorporate this information to make better predictions (for details see the description of the ML methods). Further details on the splitting strategy to reflect the described scenarios are given in the section "Model evaluation" below.

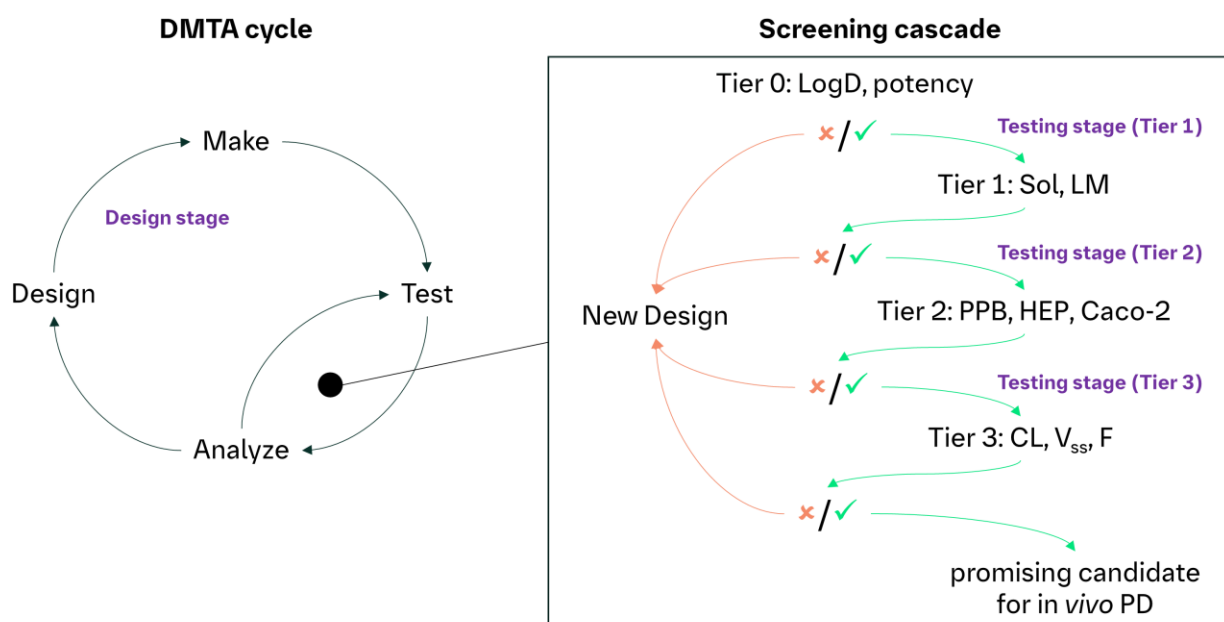


Figure 1 Evaluation scenarios in a drug discovery workflow. The design stage corresponds to virtual compounds (no experimental data available). The tiers of the screening cascade define the order in which experiments typically are conducted. At the testing stage for a certain tier, experimental data of lower tiers is available.

Data processing

Assay data was extracted from Boehringer Ingelheim's internal database along with the registration date of the compound. Censored data (i.e., with an < or > operator, corresponding to data outside the assay quantification range) was included ignoring the operator. The data was curated by considering the purity of compounds and the variance if replicate measurements were available. First, compounds with a reported purity below 80% were excluded. For measurements with replicates, the percentual coefficient of variation (%CV) was computed as a criterion for filtering. Measurements with exactly two replicates were removed if the %CV was at least 50. For measurements with more than 2 replicates, it was tested if a single clear outlier can be identified. In particular, it was tested whether removing one measurement would lead to a %CV below 50. If one outlier could be identified, this outlier was removed, and the remaining replicates were kept. In general, (remaining) replicates were aggregated by computing the arithmetic mean.

Moreover, we applied maximum filters to some of the endpoints, as results in certain ranges (very high values) were deemed unreliable. In some cases, high values were not filtered out, but instead mapped to a lower value, as the method was considered to not meaningfully distinguish values in this range. For example, some of the experimentally determined bioavailabilities (the ratio of AUC oral and AUC i.v.; AUC describes the area under the plasma concentration-time curve) were measured to be larger than 100% (e.g., due to incomplete plasma concentration-time profiles after i.v. dosing). Such measurements were set to 100% which is the highest possible bioavailability in theory. In Table 2, the maximum values for inclusion and for modeling are reported for the endpoints, where applicable. Values above the limit of inclusion were filtered out, whereas values above the limit for modeling (but below the limit for inclusion) were set to the limit for modeling.

Table 2 Limits for including data and modeling.

Endpoint	Max limit for inclusion	Max limit for modeling
rPPB	100%	99.9%
hPPB	100%	99.9%
mPPB	100%	99.9%
Caco Efflux	80	-
rCL	200 mL/(min*kg)	150 mL/(min*kg)
mCL	200 mL/(min*kg)	150 mL/(min*kg)
rVss	25 L/kg	-
mVss	25 L/kg	-
rF	150%	100%
mF	150%	100%

For some of the endpoints, transformations were applied to improve the data distribution for modeling. These were a logarithmic transformation (with base 10) for the endpoints: HTSol2.2, HTSol4.5, HTSol6.8, Caco Perm, Caco Efflux, rCL, mCL, rVss, mVss; and a logistic transformation for the endpoints: rPPB, hPPB, mPPB.

ML models

In this study we tested two single-task approaches (RF and ST-Chemprop) and two approaches considering relationships between endpoints (stacked RF and MT-Chemprop). All the trained models were regression models.

RF: A RF is an ensemble of randomized decision trees widely used for QSAR modeling. The scikit-learn implementation for regression models was used.³³ As descriptors, we used alvaDesc.³⁴ Hyperparameters were selected after initial experiments (results not shown). The number of trees in the forest was set to 500, 20 was set as the maximum depth of trees, and 50% of all features were considered when looking for the best split. Otherwise default hyperparameters from the scikit-learn version 1.1.1 were used.

Stacked RF: Model stacking, a two-step procedure, was used as a technique to leverage relationships between different endpoints. The implementation here closely follows the description in a recent publication, where this approach was referred to as Feature Net.³⁵ In the first step, a regular RF single task model is trained for each endpoint and this model is used to impute missing data for the corresponding endpoint (not all compounds have been measured in each assay). Then, a second RF model is trained for each endpoint. As features for this model, the chemical descriptors are concatenated with one column for each auxiliary endpoint. The values in the auxiliary columns are either experimentally measured, or, if not available, imputed using the models from the first step. The individual RF instances used the same hyperparameters as for single task modeling, with the exception that the models in the second step considered all available features to find the best split to ensure that relevant auxiliary features are considered.

Chemprop: The Chemprop package is a popular implementation of graph-based chemical property prediction based on the message-passing neural network (MPNN) framework. Following an initial featurization with fundamental atom and bond descriptors, learned representations of input molecules are obtained through graph convolutions. Both single task and multi-task models were trained using the Chemprop package. For early stopping, the training data was split with a scaffold-based scheme (90/10). All models were trained for up to 30 epochs and models instances stored after each epoch. In the end, the model instance with the best performance on the scaffold-based validation set was kept. All Chemprop models used are an ensemble of five individual neural network instances. This was considered a good tradeoff to increase performance compared to single instances, while limiting computational cost. In addition to the default learning hyperparameters, some variations were tested to improve single task and multi-task models. The tested sets of hyperparameters are reported in Table 3. For all other hyperparameters, no changes from the default were made. Furthermore, it was tested whether the utilization of global descriptors from RDKit could enhance the performance of the model as has been shown before for some datasets.³⁶ In the Chemprop framework, global molecular descriptors are concatenated with the learned molecular representation before passing the representation on to fully connected layers. In the SI (Table S1), representative commands for training Chemprop models are shown.

Table 3 Tested hyperparameters for Chemprop.

Set	depth	hidden_size	ffn_hidden_size
1 (default)	3	300	300
2	5	300	300
3	3	1000	1000
4	5	1000	1000
5	3	100	100

Model evaluation

All datasets were split into a training set, a validation set, and a test set according to the registration date of the molecules to evaluate the ML models in a prospective manner. In particular, the training set comprised molecules registered until the end of 2020, the validation set molecules registered in the year 2021, and the test set molecules registered in the years 2022 and 2023 (up to end of June). Note that these splits are distinct from the scaffold-based split used for early stopping. In a first step, different ML models were trained on the training set and used to predict the validation set to compare their performance. As described above, some attempts were made to further improve the performance of Chemprop models by modifying hyperparameters and adding global descriptors. Finally, the models were re-trained on both training set and validation set to predict the test set. For this we used the best model settings identified (i.e., hyperparameters and with or without global descriptors) when evaluating on the validation set.

MT-Chemprop and stacked RF were evaluated both at design stage and testing stage. At design stage, no auxiliary information about the compounds to be predicted (i.e., validation set for the first models and test set for the second models) was used during model training. For the testing stage, all available auxiliary information for lower tiers was added to the models (MT-Chemprop and stacked RF). This is illustrated in Figure 2. For MT-Chemprop the respective compounds and experimental measurements for lower tiers assays were added to the training set. For the stacked models, the experimental values were used to replace predicted values where possible.

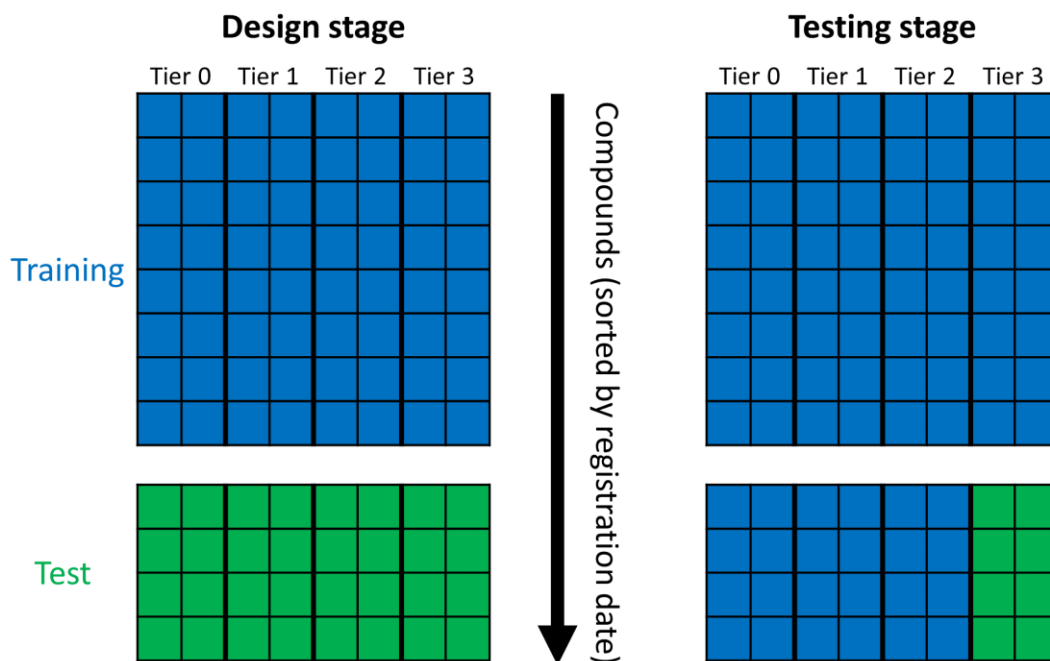


Figure 2 Visual explanation of the different evaluation scenarios. Rows correspond to compounds, while columns represent experiments (tiers separated by thick lines). At the design stage, no experimental data of test compounds is made available to the ML model. At testing stage, when available, experimental data of lower tiers than the assay evaluated is added to the training data. For example, the right panel illustrates data included for predicting *in vivo* PK characteristics (i.e., a Tier 3 study).

To evaluate the regression models, we used the coefficient of determination (R^2) and root-mean-square error (RMSE):

$$R2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i (f_i - y_i)^2} \quad (2)$$

With y_i and f_i being the (transformed) experimental and predicted value of molecule i , respectively; \bar{y} the mean (transformed) experimental value and n the number of molecules.

Biogen LM prediction

As an external validation, we used data for rat and human microsomal stability recently published (referred to as hLM_Biogen and rLM_Biogen to distinguish from in-house data).¹¹

This data was randomly split in a training (80%; hLM_Biogen: 2594 data points; rLM_Biogen:

2566 data points) and a test set (20%; hLM_Biogen: 493 data points; rLM_Biogen: 488 points). We trained single task models using only the Biogen training data (RF with AlvaDesc and ST-Chemprop) as well as MT-Chemprop models where we add the public training data to our in-house data as additional tasks. The in-house data were all data points considered within this study (i.e., training, validation and test set combined). When splitting the public data set, we ensured that overlapping compounds with our in-house dataset (i.e., identical InChI) were placed into the training set so that the MT-Chemprop model would not have seen the test compounds during training for any task.

Activity cliff analysis

For model evaluation on particularly challenging examples, activity cliff compounds (ACs) were identified in the datasets. ACs generally refer to chemically similar compounds with widely different activities and are notoriously hard to predict with ML.³⁷⁻³⁹ Each validation set compound (internal datasets), or test set compound (external datasets) was labeled as either being an AC or a non-AC. To decide whether a compound is considered an AC, the median structure activity landscape index (SALI) to its closest analogues was calculated.⁴⁰ The SALI index for a pair of compounds is defined as:

$$SALI = \frac{\Delta activity}{1 - Tanimoto} \quad (3)$$

To have activity values for different endpoints on the same scale, min-max normalization to the range 0-1 was applied to the assay data for this analysis. Tanimoto similarity was determined using Morgan fingerprints of radius 2, folded to 2048 bits from the RDKit. SALI values were calculated for a compound's five nearest neighbors, whereby only analogues with a Tanimoto similarity of at least 0.5 were considered. This was done to ensure that only sufficiently similar analogues were considered to label a compound as an AC. If the median SALI for a compound exceeds 1, a compound was labeled as an AC. The rationale behind this procedure was to identify compounds whose activity strongly differs to close analogues within clusters of compounds.

Pairwise Chemprop models

To understand the success of MT-Chemprop models at design stage, pairwise Chemprop models were trained. For three example endpoints (rHEP5, rPPB and rCl: three rat endpoints of higher tiers which strongly benefit from multi-task setting), two-task Chemprop models were trained in turn with each of the other endpoints (from the same or lower tiers) as auxiliary task. For a robust evaluation, 10 models with different random seeds were trained for each pair of endpoints. The models were trained on the training set and evaluated on the validation set. Default hyperparameters and no global descriptors were used. We analyzed how many of the training examples of the auxiliary task are not included in the training examples of the target task (i.e., number of complementary training examples). Moreover, the correlation between two endpoints was determined by computing Pearson correlation coefficients between overlapping training compounds of two endpoints. To further investigate the impact of the size of auxiliary training data, pairwise Chemprop models were also trained with the auxiliary dataset being downsampled so that all considered auxiliary training sets had the same number of complementary training compounds. Ten random samples were drawn for each endpoint pair and one pairwise model was trained for each sample.

RESULTS AND DISCUSSION

Model evaluation on the validation set

Initially, the studied ML techniques were assessed with models trained on the training set and evaluated on the validation set. As single-task models (i.e., each assay modelled separately), we investigated RF and ST-Chemprop, as models incorporating relationships between assays stacked RF and MT-Chemprop were investigated. When evaluating the latter techniques, two scenarios were distinguished: at design stage (i.e., no experimental information for predicted compounds) and at testing stage (where available, experimental information of lower tier assays is made available to models, see Material and Methods). R^2 and RMSE scores for all models on all assays are reported in Table S2 and Table S3. For both scenarios, MT-Chemprop (before optimizing the performance, see below) was clearly superior to stacked RF. Therefore, stacked

RF was not further analyzed. Figure 3 shows R^2 scores for all individual assays as well as averaged by tier for RF, ST-Chemprop, MT-Chemprop and MT-Chemprop+exp (i.e., MT-Chemprop at testing stage). Note that the numbers of validation datapoints shown in Table 1 provide an overview how much information of earlier experiments is available at testing stage.

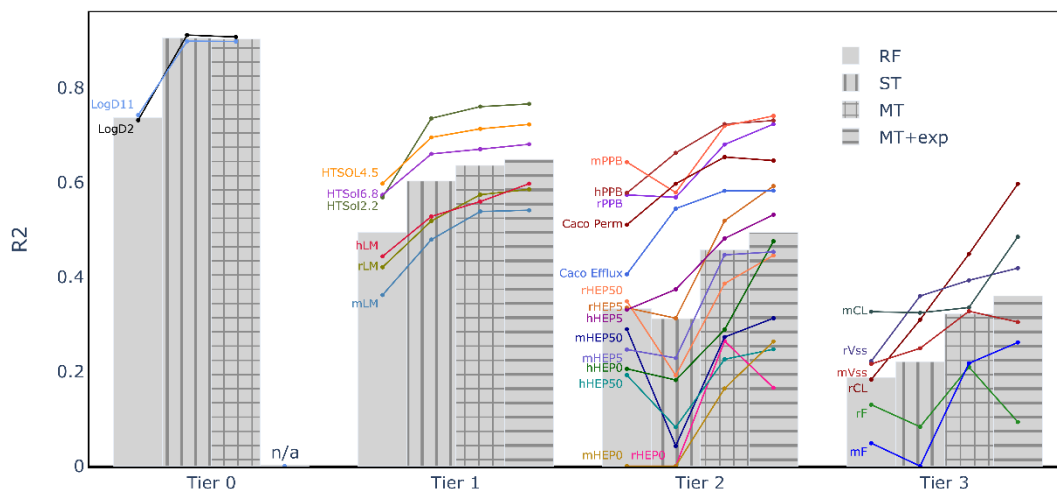


Figure 3 R^2 model scores for individual endpoints and averaged for tiers on the validation set. Lines represent individual endpoint performance, while bars represent the average of the performance for the tier.

First, clear differences between the two single-task techniques can be observed. On all Tier 0 and Tier 1 assays, RF was clearly outperformed by ST-Chemprop. These are the assays with the largest training sets, and it seems that in these cases learned representations are superior to RF with AlvaDesc as a set of conventional chemical descriptors. The performance of ST-Chemprop for Tier 2 and Tier 3 was overall comparable to RF with a tendency that assays with a high number of data points were better predicted by ST-Chemprop. This is consistent to observations in the literature that large amounts of training data are required for learned representations to outperform traditional chemical descriptors.^{36,41}

Second, MT-Chemprop used at design stage clearly outperformed both single task techniques. While MT-Chemprop provides no benefit over ST-Chemprop for the logD assays (Tier 0), it outperforms ST-Chemprop for all other assays. It seems that the presence of other tasks clearly improves the relevance of the learned representations for virtually all the assays. In a later section it is further analyzed which auxiliary tasks are most useful for the individual target tasks.

Third, an additional improvement was found when MT-Chemprop was used at the testing stage. This shows that these models can leverage experimental data of earlier assays to predict later ones more accurately for the same compounds. This modeling scenario has also been referred to as imputation (i.e., predict missing bioactivity data from partially available bioactivity data)⁴² and the benefit over traditional prediction based purely on chemical descriptors has been reported in several studies.^{35,42–45}

When considering the different tiers, model scores decrease from earlier to later tiers, regardless of what modeling technique is used. Several reasons might be contributing to that. In later tiers, fewer training data points are available to learn from for the models. Moreover, later experiments correspond to biologically more complex processes. For instance, bioavailability captures solubility, permeability, and efflux in the gut, as well as first-pass metabolism, which are all mechanisms individually tested in *in vitro* ADME assays. On the other hand, it seems that assays of later tiers very strongly benefit from the multi-task approach (i.e., larger increase in R^2 score). Small training datasets limit how well a ML model can learn structure-activity relationships. However, in a multi-task model, small tasks may benefit from the joint representation learned from all the tasks, effectively augmenting the data the model can learn from to predict small tasks.

Model tuning

In the following, we attempted to further increase the performance of the MT-Chemprop models. For this we considered modifying the learning hyperparameters as well as adding global descriptors to learn from. Model scores of all tested variants of ST-Chemprop and MT-Chemprop on the validation set can be found in the Tables S4-S7. The hyperparameter sets 2, 3 and 4 (see Material and Methods) all outperformed set 1 (i.e., the default settings), while set 5 performed worse. Among those, we selected the set 4 for all MT models, which has a larger learned representation, more neurons in the fully connected layers, as well as increased depth in the message-passing step. Also, we found that the addition of RDKit global descriptors (as implemented in the Chemprop package) generally provides a benefit in model performance. In Figure 4, we report the changes in R^2 score for all assays on the validation set with the selected setup (i.e., hyperparameter set 4 and use of RDKit descriptors) in comparison to the default scores.

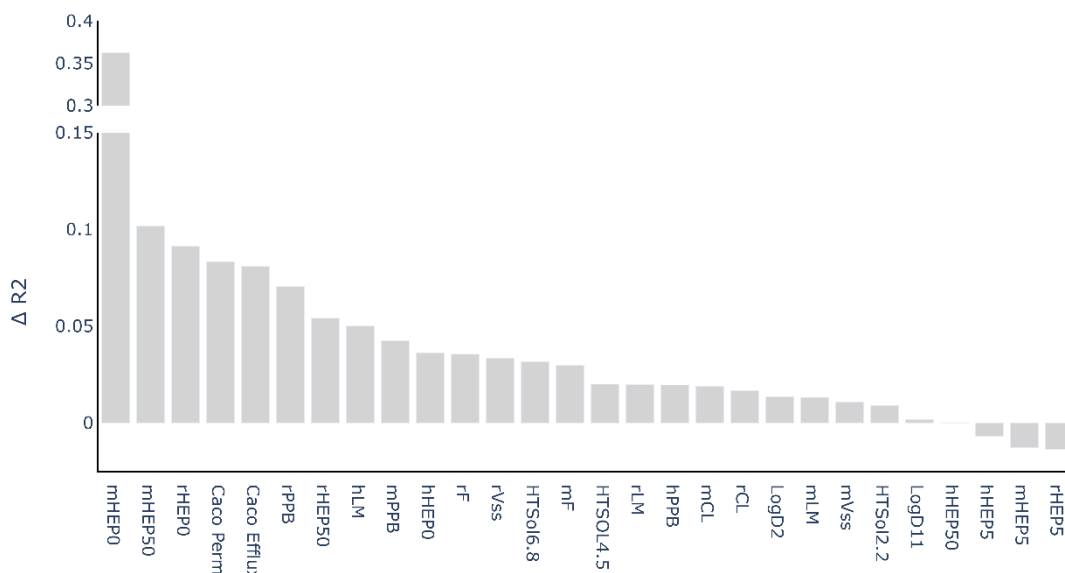


Figure 4 Improvements in R^2 score for the optimized MT-Chemprop model over the default model on the validation set.

The optimized models for most assays provide a small to moderate benefit of up to 0.1 in R^2 score. For mHEP0, an exceptionally large increase of around 0.35 was observed, yet large variations in performance for this assay may be explained by the small size of the validation set (< 100 compounds). Only for three of the endpoints a slight decrease in performance was observed. Overall, the optimized model setting provides a clear benefit over the default models (hyperparameter set 1, no RDKit descriptors) and hence was used for final evaluation on the test set. In comparison to the default models, the selected set of hyperparameters means that the model has a larger number of trainable model parameters. It appears that in our case we have sufficient training data to observe a slight increase in performance compared to the default settings. Moreover, the beneficial effect of adding global descriptors (from RDKit) indicates complementary information compared to purely learned representations from chemical graphs. As an additional control, we trained a MT-Chemprop model without graph convolutions (i.e., multi-task feedforward neural network with RDKit descriptors as input, hyperparameter set 4). This model performed worse than our optimized model on all endpoints (see Table S6 and Table S7) illustrating the importance of the representations learned by graph convolutions in the models.

Model evaluation on the test set

For the final evaluation on the test set (data has not been used for model tuning and selection), ST-Chemprop as well as MT-Chemprop (at design and testing stage) were evaluated. For a fair comparison to the optimized MT-Chemprop model, the best hyperparameter set per endpoint was used for ST-Chemprop, as well as global RDKit descriptors, if those were found beneficial on the validation set. R^2 scores for individual endpoints as well as average scores for tiers are shown in Figure 5. R^2 and RMSE scores for all models on all assays are reported in Table S8 and Table S9.

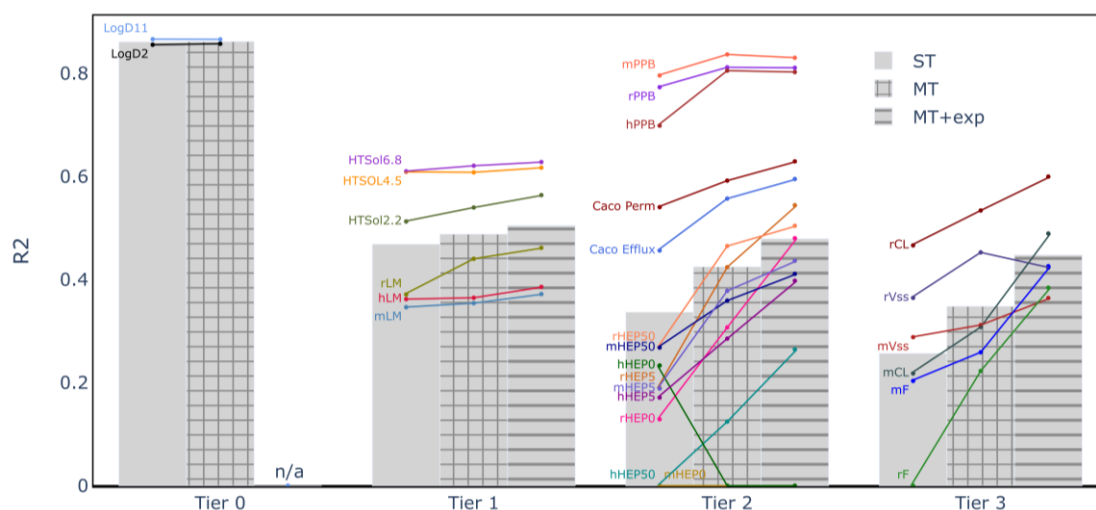


Figure 5 R^2 model scores for individual endpoints and averaged for tiers on the test set. Lines represent individual endpoint performance, while bars represent the average of the performance for the tier.

Similar as for the validation set, it can be observed that MT-Chemprop overall outperformed ST-Chemprop, with larger increases in R^2 score for the higher tiers. When evaluating MT-Chemprop at testing stage, a further benefit can be observed for most assays. As outliers to the overall trend can be identified endpoints with very small datasets (hHEP0 and mHEP0) where large variability in predictiveness can be expected. In those cases, the R^2 scores drop with MT-Chemprop (hHEP0) or are below 0 regardless of the model (mHEP0). While the same trends were observed for both evaluation steps (on validation and test set), model scores for individual endpoints may vary considerably. For instance, scores for rPPB were higher when evaluated on the test set (MT-Chemprop scores for validation/test: R^2 : 0.680/0.812; RMSE: 0.474/0.397), while the opposite was observed for hLM (MT-Chemprop scores for validation/test: R^2 : 0.559/0.365; RMSE:

15.9/16.7). Fluctuation of ML model performance over time is not uncommon. Sheridan et al found that the presence of activity cliffs (i.e., test compounds with different activities compared to similar compounds in the training set) in the test set is a key factor to explain variable model performance over time.³⁸ Nevertheless, the evaluation on the test set confirms that MT-Chemprop markedly outperformed ST-Chemprop in this analysis.

External validation

In a recent publication by researchers from Biogen, data for some ADME endpoints was made publicly available which provides an opportunity to test the benefit of MT-Chemprop models when combining in-house with public data.¹¹ For this exercise, we used the external datasets for human and rat liver microsome stability (referred to as hLM_Biogen and rLM_Biogen to distinguish from in-house data). In Figure 6A, R^2 scores of different models are reported. The models are RF (trained on Biogen data), ST-Chemprop (trained on Biogen data), MT-Chemprop (trained on hLM_Biogen and rLM_Biogen; MT-BG), and MT-Chemprop (trained in hLM_Biogen and rLM_Biogen as well as all 28 in-house endpoints; MT-BG+BI).

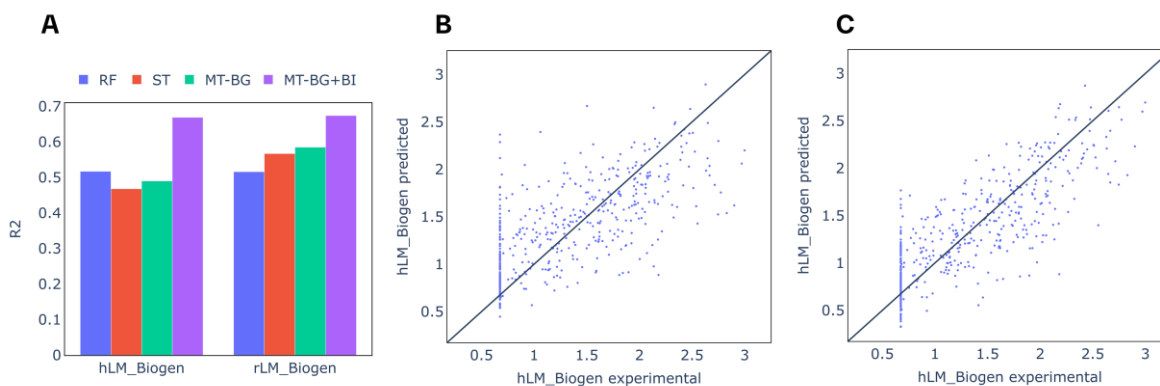


Figure 6 Model performance on the external validation set (Biogen data). **A:** R^2 scores for RF, ST-Chemprop, MT-Chemprop (BG), MT-Chemprop (BG+BI). **B:** Experimental and predicted values with identity line for ST-Chemprop on hLM. **C:** Experimental and predicted values with identity line for MT-Chemprop (BG+BI) on hLM.

It can be observed that the MT-Chemprop model including the in-house tasks clearly outperformed the other models on both hLM-Biogen and rLM-Biogen with R^2 scores of nearly 0.7. Combining the two Biogen tasks in a MT-Chemprop model outperformed the respective ST-

Chemprop models, although for hLM_Biogen, the RF model achieved a better score. Experimental and predicted values for hLM-Biogen are shown for ST-Chemprop (Figure 6B) and MT-Chemprop (BG+BI) (Figure 6C). A better correlation between experimental and predicted values can be seen for the model including the in-house tasks. These observations demonstrate how our in-house data may be used to boost the performance on external datasets. It appears that tasks with a large training set spanning a wide chemical space assist the model in predicting related tasks with less training data. We believe that this strategy may be applicable to predict many other bioactivity endpoints with limited data.

Activity cliff analysis

For a further evaluation of MT-Chemprop models, we analyzed how well the models may predict ACs, which are known to be very challenging to predict.³⁹ Hence, the objective was to determine if benefits of MT models become also apparent on ACs. Model performance on both ACs and non-ACs was evaluated for both the internal and external datasets. In Figure 7, model performances on the AC subset of the validation (internal datasets) and test sets (external datasets) are reported as relative RMSE (i.e., RMSE for the subset divided by RMSE on the full dataset for ST-Chemprop). In Figure S1, scores are also reported for the non-AC subset. As was expected, for all models relative RMSE scores are much larger for ACs in comparison to non-ACs, which indicates that ACs are predicted with lower accuracy. With respect to ACs, MT-Chemprop was the best model among the three tested for 18 out of the 22 considered endpoints with at least 20 ACs in the set used for evaluation.

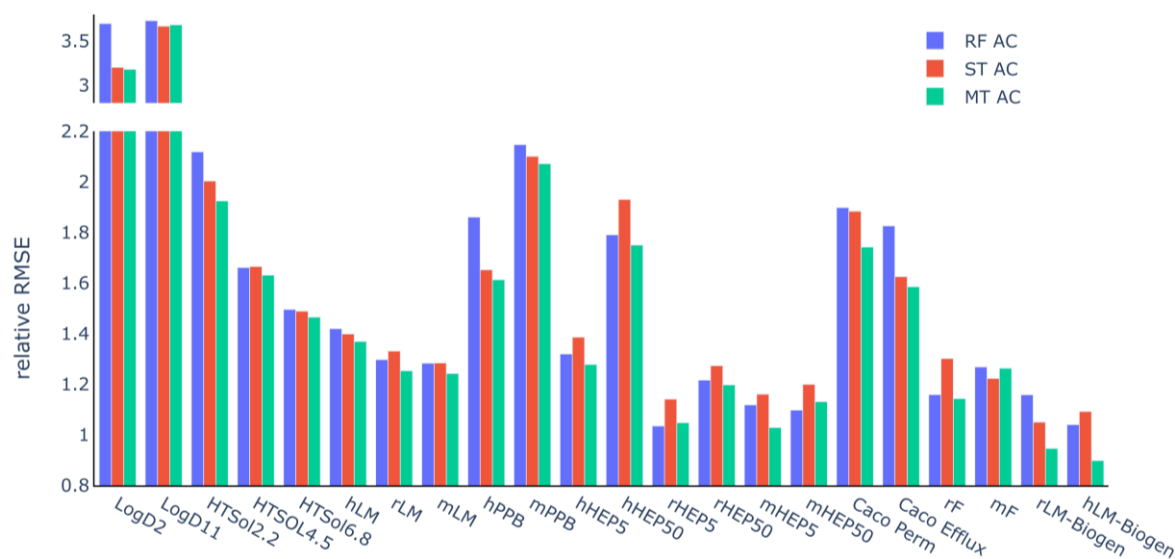
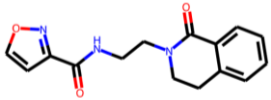
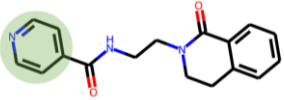
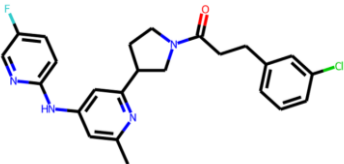
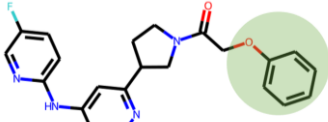
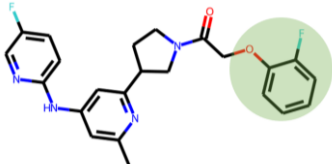


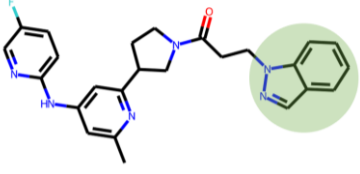
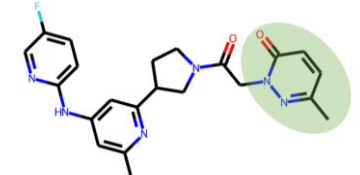
Figure 7 Models evaluated on activity cliffs (ACs). The evaluation set for each endpoint was divided in ACs and non-ACs (see Methods). Reported are RMSE scores relative to the RMSE of ST-Chemprop for the full dataset. Number of ACs per dataset: LogD2 (506), LogD11 (741), HTSol2.2 (518), HTSol4.5 (973), HTSol6.8 (1154), hLM (1319), mLM (994), hPPB (70), mPPB (23), hHEP5 (272), hHEP50 (108), rHEP5 (149), rHEP50 (112), mHEP5 (201), mHEP50 (152), Caco Perm (137), Caco Efflux (272), rF (31), mF (60), rLM-Biogen (86), hLM-Biogen (73).

Two examples in which MT-Chemprop made accurate predictions on ACs from the hLM-Biogen dataset are presented in Table 4. On average, errors for MT-Chemprop are around 14% smaller than for RF and 18% smaller than for ST-Chemprop for this task (numbers obtained by dividing MT-Chemprop RMSE for ACs by the single task model RMSE for ACs). Mol2977 is the only close analogue of Mol2062 with a Tanimoto similarity of at least 0.5 (pyridine ring instead of isoxazole) and the two compounds differ moderately in stability (13.6 mL/min/kg vs 4.74 mL/min/kg) and hence Mol2062 was labelled as AC (see Material and Methods). MT-Chemprop's and RF's predictions for this AC are much closer to the experiment (5.43 and 5.37 mL/min/kg, respectively) than the prediction from ST-Chemprop (25.4 mL/min/kg). In the second example, four close analogues of Mol1903 were identified. Both Mol1903 and Mol875 were moderately or very well predicted by MT-Chemprop, respectively, whereas both single task techniques clearly underpredicted the clearance for both compounds. Those examples illustrate how MT-Chemprop seems to leverage information from auxiliary tasks to accurately predict the microsomal clearance in the presence of ACs. While ACs are still challenging to predict for MT-

Chemprop (larger errors compared to non-ACs), it appears that the overall observed benefit over the single task models also holds true for ACs. Previously, it was shown that several graph-based neural networks are inferior to traditional ML models (such as SVM) when predicting ACs.³⁹ However, our results suggest that graph-based models such as Chemprop may be a good choice to predict also ACs, provided that the endpoint of interest can benefit from a multi-task setting.

Table 4 Example predictions for ACs and their close analogues for hLM-Biogen^a

	Compound	Exp hLM-Biogen	Pred (RF / ST) hLM-Biogen	Pred (MT) hLM-Biogen
Example 1	 Mol2062 - query	4.74	5.37 / 25.4	5.43
	 Mol2977	13.6	Training	Training
Example 2	 Mol1903 - query	562	107 / 199	441
	 Mol1979	106	Training	Training
	 Mol875	115	69.7 / 63.9	117

 <p>Mol2207</p>	345	Training	Training
 <p>Mol2500</p>	19.7	Training	Training

^aShown are two example query compounds found to be ACs and their nearest neighbors (Tanimoto >0.5). The columns show experimental data as well as predictions by RF, ST-Chemprop and MT-Chemprop, in case the compound was in the test set. Experimental and predicted clearances have the unit mL/min/kg.

Analysis on auxiliary task relevance

Having established a superior performance of MT-Chemprop models to ST models, we sought to understand the success of those models focusing on evaluations at design stage (see Figures 1 and 2). We attempted to attribute the success to certain auxiliary tasks. For that purpose, pairwise (i.e., two-task) Chemprop models were trained for three exemplary target endpoints: rHEP5, rPPB and rCL. In Figure 8 A-C, the performance of each pairwise model is contrasted against the size of the respective auxiliary task, and in D-F against the Pearson correlation of overlapping compounds in the training set as a measure of task relatedness.

The rHEP5 task strongly benefitted from the physiologically related LM tasks (also reflected in a moderate to strong Pearson correlation). Similarly, rLM as auxiliary task strongly improved the predictions for rCL. The rHEP assays are even more closely related to *in vivo* clearance, yet for those endpoints much less training data was available which might have prevented a benefit on accuracy. Neither were the predictions for rPPB improved when the strongly correlated hPPB or mPPB endpoints were added as tasks. On the other hand, moderately correlated endpoints like LogD11 (positive correlation) or HTSol 6.8 (negative correlation) with large amounts of training data were successfully used as auxiliary tasks.

Overall, the large auxiliary tasks were found most useful as auxiliary tasks in MT-Chemprop models, despite only weak or moderate correlation to the target assays. Interestingly, simple experiments like LogD2 and LogD11 that measure lipophilicity can be successfully leveraged as auxiliary tasks. Trends between lipophilicity and PK parameters are well established in the literature which may explain our findings.⁴⁶

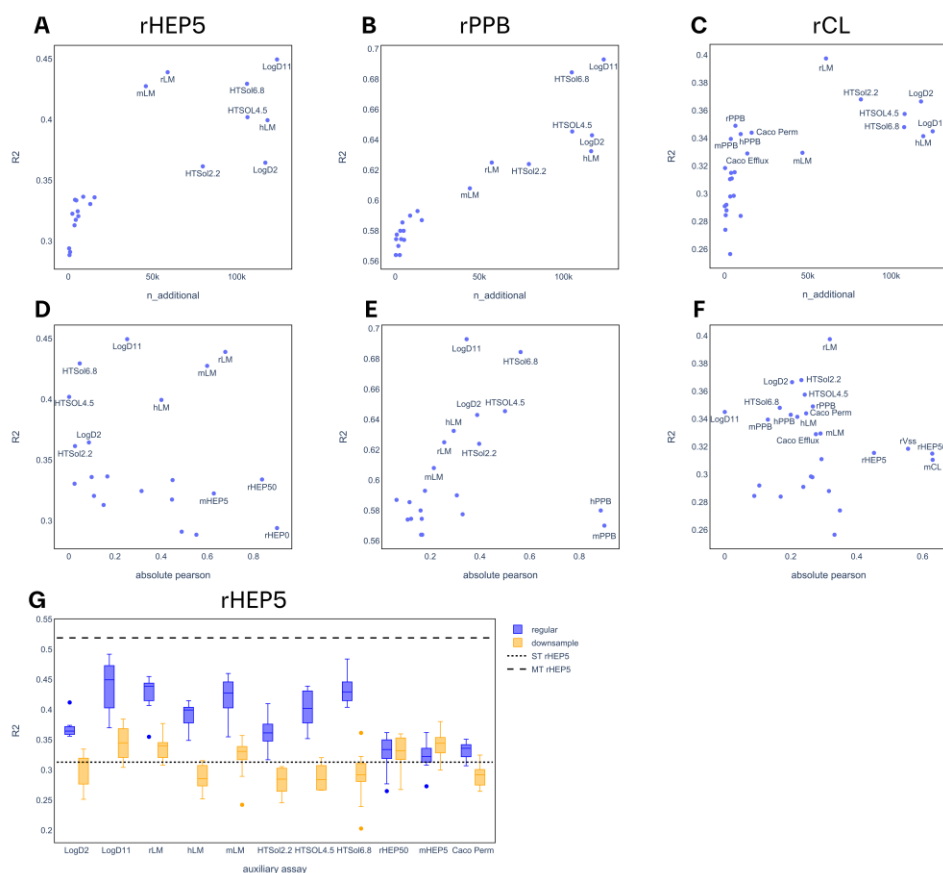


Figure 8 Analysis of pairwise Chemprop models for rHEP5, rPPB and rCL. **A-C:** Median R^2 scores of pairwise Chemprop models versus size of the auxiliary task (i.e., number of complementary training compounds). **D-F:** Median R^2 scores of pairwise Chemprop models versus Pearson correlation of overlapping compounds between target and auxiliary task. **G:** R^2 scores of pairwise Chemprop models for rHEP5 as target (10 random seeds) in comparison to R^2 scores if auxiliary task is downsampled to smallest considered auxiliary task (here mHEP5).

To further understand the role of auxiliary task size, we also downsampled the auxiliary tasks so that all compared auxiliary task are of same size for rHEP5 (Figure 8G). This resulted in clear

decreases in performance for all pairwise models with downsampled auxiliary tasks. Comparable observations can be made for rPPB and rCL (see Figure S2). These findings confirm the importance of the size of auxiliary tasks for the success of MT-Chemprop model in our case.

In a previous study on benefits of multi-task models over single-task models, correlation between target and auxiliary was found as a prerequisite to achieve an improvement over a single-task model.⁴⁷ Our results suggest that even a relatively weak correlation (e.g., Pearson correlation coefficient between rHEP5 and LogD11 ~0.2) may be sufficient for a strong multi-task benefit. The benefit of large auxiliary tasks may be attributed to a larger chemical space that the model encompasses. Our findings are consistent with observations from other studies. For instance, the prediction of *in vivo* brain penetration could be strongly improved when learned together with much larger *in vitro* tasks related to lipophilicity and membrane permeation.²⁹ Prediction of aqueous solubility from powder material as a relatively small task was improved when trained together with other physicochemical property tasks such as predicting solubility measured with other methods and LogD measured at different pH.¹⁵ These findings can be translated into a general strategy for endpoints with little training data. By combining them with much larger auxiliary tasks that are at least weakly correlated in a multi-task model, improved predictions can be achieved in many cases.

CONCLUSIONS

In the present study, we demonstrated the usefulness of multi-task modeling for predicting ADME/PK properties using in-house data from drug discovery projects at Boehringer Ingelheim. Graph-based multitask learning proved to be useful to train the models. Predicting ADME/PK parameters at industrial scales could be identified as a use case where learned representations from chemical graphs appear advantageous in comparison to traditional chemical descriptors. Using a multi-task setting means having a joint representation of compounds across tasks which benefitted especially smaller tasks (complex *in vitro* endpoints or *in vivo* PK). We have shown that for ADME properties, a single MT model performs well, even beyond previously reported sets of endpoints.⁴¹

The graph-based multitask models could be successfully extended to external data. By combining in-house tasks with external ones (hLM and rLM) from a public data set, performance on external tasks could be clearly improved in comparison to single-task or multi-task models based purely on the external data. In addition to improved performance, a multi-task model means that the number of ML models that need to be deployed and maintained for productive use can be reduced. Regular model updates ensure that currently researched chemical space is covered and that information about early assays can support predictions for more complex assays (i.e., predictions at testing stage).

While relatively well explored for ADME predictions, more work is needed to fully understand the potential of multi-task modeling for other bioactivity tasks such as potency or toxicity prediction. Future studies could investigate whether physicochemical datasets such as LogD or aqueous solubility as auxiliary tasks could also improve predictivity of other types of endpoints. A concept related to multi-task learning is transfer learning, where a neural network is pre-trained on a large dataset before fine-tuning the model using data from the target task.^{16–18,48,49} A direction for further interesting research will be to systematically understand under which circumstances multi-task or transfer learning strategies are superior.

ASSOCIATED CONTENT

Data and software availability statement

The in-house data used for this study is confidential. The external benchmark dataset released by Biogen used in this study (including train-test split used) is available in the Supporting Information. All models were trained using modeling tools in the public domain (Chemprop and scikit-learn for RF) and the used hyperparameters are shared. Exemplary commands to train Chemprop models are shown in Table S1. The code used to label compounds as ACs is shared as a Jupyter Notebook (see Supporting Information).

Supporting Information

Supporting_Information.pdf: Table S1: exemplary commands to train a Chemprop model and make predictions. Tables S2-S9: R^2 and RMSE scores of models on individual endpoints. Figure S1: Model performances on ACs and non-ACs. Figure S2: R^2 scores of downsampled pairwise MT-Chemprop models for the endpoints rPPB and rCL.

Supporting_Material.zip: Training data used to train models for hLM_Biogen and rLM_Biogen. (Biogen_train_hlm_rlm.csv)

Test data used to evaluate models for hLM_Biogen and rLM_Biogen.

(Biogen_test_hlm_rlm.csv)

Code to demonstrate labeling of compounds as ACs. (activity_cliff_labeling.ipynb)

Code to demonstrate labeling of compounds as ACs. (activity_cliff_labelling.html)

Author Contributions

Conceptualization: M.W., L.H., M.S. Data collection and curation: M.W., L.H., M.S.

Development and Implementation of ML models: M.W. Data analysis: M.W. Writing: M.W.

J.M.B., L.H., M.S.

Notes

The authors declare no competing financial interest.

Acknowledgement

We would like to thank the groups and teams at Boehringer Ingelheim who generated the data in this study, namely Research PK, *In vitro* ADME, Bioanalysis, and CMC as part of the Drug Discovery Sciences Department. We further wish to thank Dr. Christofer Tautermann, Dr. Nils Weskamp, Prof. Dr. J. B. Brown, Dr. Alexander Weber, Dr. Igor Zingman for constructive feedback during the preparation of the manuscript.

Abbreviations

AC, activity cliff; ADME, absorption distribution metabolism excretion; AUC, area under the plasma concentration-time curve; ML, machine learning; MPNN, message-passing neural network; PK, pharmacokinetics; QSAR, quantitative structure-activity relationship; QSPR, quantitative structure-property relationship; R^2 , coefficient of determination; RF, random forest; RMSE, root mean square error; SALI, structure-activity landscape index; SVM, support vector machine.

References

- (1) Obrezanova, O. Artificial Intelligence for Compound Pharmacokinetics Prediction. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102546. <https://doi.org/10.1016/j.sbi.2023.102546>.
- (2) Danishuddin; Kumar, V.; Faheem, M.; Lee, K. W. A Decade of Machine Learning-Based Predictive Models for Human Pharmacokinetics: Advances and Challenges. *Drug Discov. Today* **2022**, *27* (2), 529–537. <https://doi.org/10.1016/j.drudis.2021.09.013>.
- (3) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010. <https://doi.org/10.1021/jm4004285>.
- (4) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. <https://doi.org/10.1039/c8sc00148k>.
- (5) Robinson, M. C.; Glen, R. C.; Lee, A. A. Validating the Validation: Reanalyzing a Large-Scale Comparison of Deep Learning and Machine Learning Models for Bioactivity Prediction. *J. Comput.-Aided Mol. Des.* **2020**, *34* (7), 717–730. <https://doi.org/10.1007/s10822-019-00274-0>.
- (6) Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; Wang, F. A Systematic Study of Key Elements Underlying Molecular Property Prediction. *Nat. Commun.* **2023**, *14* (1), 6395. <https://doi.org/10.1038/s41467-023-41948-6>.

- (7) Aleksić, S.; Seeliger, D.; Brown, J. B. ADMET Predictability at Boehringer Ingelheim: State-of-the-Art, and Do Bigger Datasets or Algorithms Make a Difference? *Mol. Inform.* **2022**, *41* (2), e2100113. <https://doi.org/10.1002/minf.202100113>.
- (8) Kumar, K.; Chupakhin, V.; Vos, A.; Morrison, D.; Rassokhin, D.; Dellwo, M. J.; McCormick, K.; Paternoster, E.; Ceulemans, H.; DesJarlais, R. L. Development and Implementation of an Enterprise-Wide Predictive Model for Early Absorption, Distribution, Metabolism and Excretion Properties. *Futur. Med. Chem.* **2021**, *13* (19), 1639–1654. <https://doi.org/10.4155/fmc-2021-0138>.
- (9) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET Prediction with Multitask Deep Featurization. *J. Med. Chem.* **2020**, *63* (16), 8835–8848. <https://doi.org/10.1021/acs.jmedchem.9b02187>.
- (10) Rodríguez-Pérez, R.; Trunzer, M.; Schneider, N.; Faller, B.; Gerebtzoff, G. Multispecies Machine Learning Predictions of In Vitro Intrinsic Clearance with Uncertainty Quantification Analyses. *Mol. Pharm.* **2023**, *20* (1), 383–394. <https://doi.org/10.1021/acs.molpharmaceut.2c00680>.
- (11) Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, C.; Trapa, P.; Sciabola, S. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. *J. Chem. Inf. Model.* **2023**, *63* (11), 3263–3274. <https://doi.org/10.1021/acs.jcim.3c00160>.
- (12) Kearnes, S.; Goldman, B.; Pande, V. Modeling Industrial ADMET Data with Multitask Networks. *arXiv* **2016**. <https://doi.org/10.48550/arxiv.1606.08793>.
- (13) Mora, A. M.; Subramanian, V.; Miljković, F. Multi-Task Convolutional Neural Networks for Predicting in Vitro Clearance Endpoints from Molecular Images. *J. Comput.-Aided Mol. Des.* **2022**, *36* (6), 443–457. <https://doi.org/10.1007/s10822-022-00458-1>.
- (14) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (3), 1253–1268. <https://doi.org/10.1021/acs.jcim.8b00785>.
- (15) Montanari, F.; Kuhnke, L.; Laak, A. T.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2019**, *25* (1), 44. <https://doi.org/10.3390/molecules25010044>.
- (16) Stoyanova, R.; Katzberger, P. M.; Komissarov, L.; Khadhraoui, A.; Sach-Peltason, L.; Zbinden, K. G.; Schindler, T.; Manevski, N. Computational Predictions of Nonclinical Pharmacokinetics at the Drug Design Stage. *J. Chem. Inf. Model.* **2023**, *63* (2), 442–458. <https://doi.org/10.1021/acs.jcim.2c01134>.
- (17) Schneckener, S.; Grimbs, S.; Hey, J.; Menz, S.; Osmers, M.; Schaper, S.; Hillisch, A.; Göller, A. H. Prediction of Oral Bioavailability in Rats: Transferring Insights from in Vitro

Correlations to (Deep) Machine Learning Models Using in Silico Model Outputs and Chemical Structure Parameters. *J. Chem. Inf. Model.* **2019**, *59* (11), 4893–4905.

<https://doi.org/10.1021/acs.jcim.9b00460>.

(18) Ye, Z.; Yang, Y.; Li, X.; Cao, D.; Ouyang, D. An Integrated Transfer Learning and Multitask Learning Approach for Pharmacokinetic Parameter Prediction. *Mol. Pharm.* **2019**, *16* (2), 533–541. <https://doi.org/10.1021/acs.molpharmaceut.8b00816>.

(19) Miljković, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. *Mol. Pharm.* **2021**, *18* (12), 4520–4530.

<https://doi.org/10.1021/acs.molpharmaceut.1c00718>.

(20) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* *50* (5), 742–754. <https://doi.org/10.1021/ci100050t>.

(21) Yap, C. W. PaDEL-descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474.

<https://doi.org/10.1002/jcc.21707>.

(22) Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *Mol. Inform.* *20* (3), 227–240.

[https://doi.org/10.1002/1521-3838\(200110\)20:3<;227::aid-qsar227>3.0.co;2-y](https://doi.org/10.1002/1521-3838(200110)20:3<;227::aid-qsar227>3.0.co;2-y).

(23) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958. <https://doi.org/10.1021/ci034160g>.

(24) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45* (3), 786–799. <https://doi.org/10.1021/ci0500379>.

(25) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274.

<https://doi.org/10.1021/ci500747n>.

(26) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019.

(27) Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; Karypis, G. DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science. *ACS Omega* **2021**, *6* (41), 27233–27238. <https://doi.org/10.1021/acsomega.1c04017>.

(28) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2023**. <https://doi.org/10.1021/acs.jcim.3c01250>.

- (29) Hamzic, S.; Lewis, R.; Desrayaud, S.; Soylu, C.; Fortunato, M.; Gerebtzoff, G.; Rodríguez-Pérez, R. Predicting In Vivo Compound Brain Penetration Using Multi-Task Graph Neural Networks. *J. Chem. Inf. Model.* **2022**, *62* (13), 3180–3190. <https://doi.org/10.1021/acs.jcim.2c00412>.
- (30) Obrezanova, O.; Martinsson, A.; Whitehead, T.; Mahmoud, S.; Bender, A.; Miljković, F.; Grabowski, P.; Irwin, B.; Oprisiu, I.; Conduit, G.; Segall, M.; Smith, G. F.; Williamson, B.; Winiwarter, S.; Greene, N. Prediction of In Vivo Pharmacokinetic Parameters and Time–Exposure Curves in Rats Using Machine Learning from the Chemical Structure. *Mol. Pharm.* **2022**, *19* (5), 1488–1504. <https://doi.org/10.1021/acs.molpharmaceut.2c00027>.
- (31) Bender, A.; Cortes-Ciriano, I. Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discov. Today* **2021**, *26* (4), 1040–1052. <https://doi.org/10.1016/j.drudis.2020.11.037>.
- (32) Caruana, R. Multitask Learning. *Mach. Learn.* **1997**, *28* (1), 41–75. <https://doi.org/10.1023/a:1007379606734>.
- (33) Pedregosa, F.; Varoquax, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *JMLR* **2011**, *12*, 2825–2830. <https://doi.org/10.48550/arxiv.1201.0490>.
- (34) Mauri, A. AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. **2020**, 801–820. https://doi.org/10.1007/978-1-0716-0150-1_32.
- (35) Walter, M.; Allen, L. N.; León, A. de la V. de; Webb, S. J.; Gillet, V. J. Analysis of the Benefits of Imputation Models over Traditional QSAR Models for Toxicity Prediction. *J. Cheminformatics* **2022**, *14* (1), 32. <https://doi.org/10.1186/s13321-022-00611-w>.
- (36) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388. <https://doi.org/10.1021/acs.jcim.9b00237>.
- (37) Maggiora, G. M. On Outliers and Activity Cliffs-Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535–1535. <https://doi.org/10.1021/ci060117s>.
- (38) Sheridan, R. P.; Culberson, J. C.; Joshi, E.; Tudor, M.; Karnachi, P. Prediction Accuracy of Production ADMET Models as a Function of Version: Activity Cliffs Rule. *J. Chem. Inf. Model.* **2022**, *62* (14), 3275–3280. <https://doi.org/10.1021/acs.jcim.2c00699>.
- (39) Tilborg, D. van; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951. <https://doi.org/10.1021/acs.jcim.2c01073>.

- (40) Guha, R.; Drie, J. H. V. Structure–Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48* (3), 646–658. <https://doi.org/10.1021/ci7004093>.
- (41) Göller, A. H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; Laak, A. ter; Wichard, J.; Lobell, M.; Hillisch, A. Bayer’s in Silico ADMET Platform: A Journey of Machine Learning over the Past Two Decades. *Drug Discov. Today* **2020**, *25* (9), 1702–1709. <https://doi.org/10.1016/j.drudis.2020.07.001>.
- (42) Irwin, B. W. J.; Mahmoud, S.; Whitehead, T. M.; Conduit, G. J.; Segall, M. D. Imputation versus Prediction: Applications in Machine Learning for Drug Discovery. *Futur. Drug Discov.* **2020**, *2* (2), FDD38. <https://doi.org/10.4155/fdd-2020-0008>.
- (43) Sosnina, E. A.; Sosnin, S.; Fedorov, M. V. Improvement of Multi-Task Learning by Data Enrichment: Application for Drug Discovery. *J. Comput.-Aided Mol. Des.* **2023**, *37* (4), 183–200. <https://doi.org/10.1007/s10822-023-00500-w>.
- (44) Irwin, B. W. J.; Levell, J. R.; Whitehead, T. M.; Segall, M. D.; Conduit, G. J. Practical Applications of Deep Learning To Impute Heterogeneous Drug Discovery Data. *J. Chem. Inf. Model.* **2020**, *60* (6), 2848–2857. <https://doi.org/10.1021/acs.jcim.0c00443>.
- (45) Whitehead, T. M.; Strickland, J.; Conduit, G. J.; Borrel, A.; Mucs, D.; Baskerville-Abraham, I. Quantifying the Benefits of Imputation over QSAR Methods in Toxicology Data Modeling. *J. Chem. Inf. Model.* **2023**. <https://doi.org/10.1021/acs.jcim.3c01695>.
- (46) Lombardo, F.; Berellini, G.; Obach, R. S. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug Metab. Dispos.* **2018**, *46* (11), dmd.118.082966. <https://doi.org/10.1124/dmd.118.082966>.
- (47) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57* (10), 2490–2504. <https://doi.org/10.1021/acs.jcim.7b00087>.
- (48) Ng, S. S. S.; Lu, Y. Evaluating the Use of Graph Neural Networks and Transfer Learning for Oral Bioavailability Prediction. *J. Chem. Inf. Model.* **2023**, *63* (16), 5035–5044. <https://doi.org/10.1021/acs.jcim.3c00554>.
- (49) Liu, R.; Laxminarayan, S.; Reifman, J.; Wallqvist, A. Enabling Data-Limited Chemical Bioactivity Predictions through Deep Neural Network Transfer Learning. *J. Comput.-Aided Mol. Des.* **2022**, *36* (12), 867–878. <https://doi.org/10.1007/s10822-022-00486-x>.