JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Predicting In Vivo Compound Brain Penetration Using Multi-task Graph Neural Networks

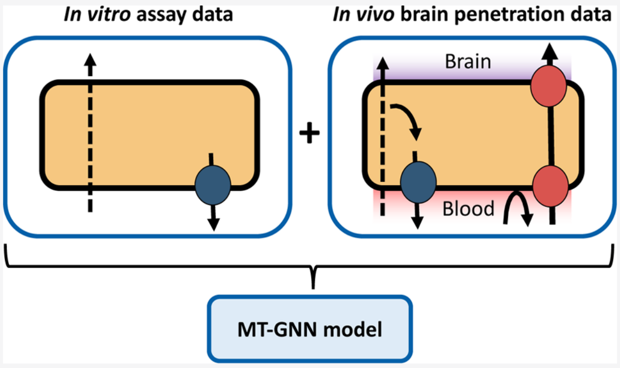
Seid Hamzic, Richard Lewis, Sandrine Desrayaud, Cihan Soylu, Mike Fortunato, Grégori Gerebtzoff, and Raquel Rodríguez-Pérez*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 3180−3190

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Assessing whether compounds penetrate the brain can become critical in drug discovery, either to prevent adverse events or to reach the biological target. Generally, pre-clinical in vivo studies measuring the ratio of brain and blood concentrations (K_p) are required to estimate the brain penetration potential of a new drug entity. In this work, we developed machine learning models to predict in vivo compound brain penetration (as LogK_p) from chemical structure. Our results show the benefit of including in vitro experimental data as auxiliary tasks in multi-task graph neural network (MT-GNN) models. MT-GNNs outperformed single-task (ST) models solely trained on in vivo brain penetration data. The best-performing MT-GNN regression model achieved a coefficient of determination of 0.42 and a mean absolute error of 0.39 (2.5-fold) on a prospective validation set and outperformed all tested ST models. To facilitate decision-making, compounds were classified into brain-penetrant or non-penetrant, achieving a Matthew's correlation coefficient of 0.66. Taken together, our findings indicate that the inclusion of in vitro assay data as MT-GNN auxiliary tasks improves in vivo brain penetration predictions and prospective compound prioritization.



INTRODUCTION

The blood−brain barrier (BBB) is a complex biological border composed of various cell types which regulates the transport of molecules from the blood to the central nervous system (CNS) and vice versa. The physiological role of the BBB is to maintain a stable microenvironment and ensure neural function and protection from CNS-damaging compounds.[1] Drug delivery through BBB constitutes the main route to reach CNS targets,[1,2] but the high BBB complexity and selectivity makes the design of brain-penetrant compounds a challenging task.[2] However, undesired brain penetration can also occur when compounds not targeting the CNS penetrate the brain and lead to adverse effects. Thus, brain penetration can be a difficult hurdle to overcome in early drug development.[2,3]

Pre-clinical in vivo studies are required to assess brain penetration. Generally, the compound is administered, and its concentration in brain and blood is measured at certain time points. The ratio of total or free brain/blood concentrations (K_p and K_{puu}, respectively) is used as a surrogate marker for brain penetration.[4] Arguably, the free ratio K_{puu} gives a better estimation of brain penetration than K_p but requires additional experiments to estimate plasma protein and brain binding.[5] Brain penetration assessment studies involve animal testing and are not suited for large screens. To facilitate informed decisions for early compound prioritization, in vitro assays are applied for routine screening in pharmaceutical industry, for example, in vitro monolayer-based assays.[4,6] Moreover, several in silico methods have been developed, ranging from simple cut-off guidelines based on physicochemical properties to more sophisticated predictive models.[7] Since K_{puu} data is very scarce, the ratio based on total concentrations or K_p is typically utilized for in silico modeling. Rules based on calculated compound properties were initially proposed to guide the design of brain-penetrant compounds[3,8] but were typically restrictive of the design space since only a reduced set of compounds fulfilled all criteria. Multi-parameter optimization (MPO) scores aim at balancing multiple parameters to achieve improved ranking of CNS compounds.[9,10] The "CNS-MPO"[9] and, the more recently published, the "BBB-score"[10] constitute relevant examples.

Recent progress in machine learning (ML) made it possible to model brain penetration using a larger number of molecular descriptors or fingerprints. Since most publicly available brain

Received: April 9, 2022
Published: June 23, 2022

https://doi.org/10.1021/acs.jcim.2c00412

penetration data is categorical, literature is directed toward classification models that discriminate between brain-penetrant (BP+) and non-brain-penetrant (BP−) compounds. For such classification, a K_p value of 0.1 is commonly used as the threshold to classify compounds as brain-penetrant ($K_p > 0.1$) or non-penetrant ($K_p \leq 0.1$).[4,11−18] This definition typically leads to imbalanced data sets dominated by BP+ compounds,[12−14,19,20] which imposes further modeling challenges and requires the use of adequate performance metrics, such as Matthew's correlation coefficient (MCC).[21] Regression models were less common and usually based on reduced data sets.[14,16,19,22]

So far, publicly available data was mainly used for the development and evaluation of brain penetration models.[12−14,16,19] Despite some valuable efforts toward data set aggregation and standardization,[20] brain penetration data sets still remain heterogeneous and comparatively small for ML approaches, ranging from few hundreds to few thousands of molecules.[12−14,16,19,20] The most recently published data set "B3DB" constitutes the largest publicly available in vivo brain penetration data set[11] and, to the best of our knowledge, it has not been used for modeling yet. Perhaps, due to the limited data set size, previous studies have mainly reported on model performance on random compound subsets,[12,16−18,22] which is an indicator of self-consistency but not of future model predictivity.[15,23] For a more realistic estimation of model prospective performance, evaluation should be done on new chemical series or scaffolds (series or scaffold split)[15,20] or on the most recent experiments (temporal split). The latter resembles the model use in pharmaceutical research and requires temporal or date information, which is typically only available in proprietary data sets.[15,23]

Despite the recent popularity of deep learning methodologies, earlier studies applied single-task (ST) models; that is, models were solely trained on brain penetration data and using pre-computed descriptors. However, simultaneous training of multiple related endpoints, the so-called multi-task (MT) learning, has shown the potential to improve model performance on endpoints with limited amounts of training data.[24−26] Furthermore, larger data sets enable the use of graph neural networks (GNNs), which learn task-specific representations without relying on pre-computed descriptors.[15,27] Despite the recent advent of MT learning and GNNs, such methodologies are yet to be evaluated for in vivo brain penetration predictions.

In this work, we developed standard ML and GNN regression models for the prediction of in vivo brain permeation relying both on proprietary data and the largest publicly available brain penetration data set. We investigated the potential benefit of MT learning, including data from in vitro assays. Models were validated on new chemical series (scaffold split) and prospectively (temporal split) to have a realistic performance estimation and mimic the conditions of future model usage. Last, we also investigated classification approaches and compared them to classical rule-based cut-offs and MPO.

■ MATERIALS AND METHODS

Calculation Outline. Different experiments were carried out to address the following questions:

(i) Is it possible to develop a ML model to predict in vivo brain penetration from compound structure? A variety of ML models were evaluated on their ability to predict compound brain penetration in realistic settings (e.g., scaffold or temporal split).

(ii) How is model quality affected by using different data sources? Models were built and validated with the literature and proprietary data.

(iii) Can we leverage novel strategies such as MT graph neural networks (MT-GNNs) to improve compound brain penetration predictions? MT-GNN models were generated and benchmarked against more standard methods, namely, ST ML.

(iv) Which auxiliary tasks are informative for brain penetration prediction using MT-GNNs? MT-GNNs were built with different tasks and varying training set size.

Data Curation. *Novartis Internal Data Set.* From the Novartis internal database, we retrieved ∼180,000 compounds registered prior to 2019 and with the experimental data for 12 properties related to brain penetration. Table S1 lists the included assays, which measure octanol/water partition and distribution coefficients (logP/D), in vitro permeability, active efflux, and in vivo brain penetration (K_p). Table 1 reports the assembled data sets and tasks included in our data set for modeling. All experimental values except logP/D were log-transformed, and outliers were removed. The K_p data set was limited to wild-type mouse and rat with no inhibitor treatment for P-glycoprotein (P-gp). Since brain exposure is normalized by the corresponding systemic exposure, the route of administration does not influence the ratio between the brain concentration and the circulating blood concentration. Therefore, compounds administered intravenously with a dose ≤ 5 mg/kg and orally with a dose ≤ 30 mg/kg were kept. Furthermore, K_p values based on sampling time points earlier

Table 1. Novartis Internal Data set Description[a]

assay tasks	# CPDs	included in the six-task model	included in the 12-task model
HT PAMPA permeability	137,857	X	X
passive permeability LE-MDCK Papp a to b (assay version 1)	33,937	X	X
passive permeability Caco-2 Papp b to a	31,487		X
passive permeability Caco-2 Papp a to b	30,936		X
passive permeability Caco-2 efflux ratio	30,763	X	X
Log*D* in octanol buffer	20,495		X
Log*P* in octanol buffer	14,476		X
MDCK-MDR1 permeability b to a	6306		X
MDCK-MDR1 permeability a to b	6266		X
MDCK-MDR1 efflux ratio	6248	X	X
in vivo brain penetration (K_p)	2103	X	X
passive permeability LE-MDCK Papp a to b (assay version 2)	985	X	X

[a]Reported are the assay tasks included in the 6-task and 12-task models (X), and the number of compounds per task (# CPDs). The 6-task model includes the primary assay endpoints, whereas the 12-task model includes 6 additional endpoints measured in the in vitro assays. More assay information is shown in Table S1.

than 5 min or exceeding 8 h were removed. Log transformation of K_p (LogK_p) values was applied due to the non-normal distribution of the data and dynamic range of the assay. Correlations between K_p values and assay measurements are reported in Table S2. For prospective model evaluation, brain penetration data was retrieved for 111 additional compounds registered between 2019 and April 2021.

Literature Data Set. We used the currently largest publicly available brain penetration data set, known as "B3DB" from Meng et al.[11] It includes widely adopted data sets such as Wang et al.[12] and Brito-Sánchez et al.[14] augmented with additional sources. The raw data set contains classified data for 7807 compounds, as well as the continuous value for 1058 of them (13.6%). Numerical values were considered for the development of regression models. Other data sets were assembled to use as auxiliary tasks for MT learning and are reported in Table 2. Molecules were standardized using RDKit,[28] and duplicates were removed. Caco-2 apparent permeability (Papp) and MDR1 efflux ratio were log-transformed. Esposito et al.[29] also reported a coefficient of variation if multiple values were available for the same compound. We removed compounds with a coefficient of variation $\geq$ 0.5. For each task, experimental values within three standard deviations of the internal data set mean were kept, whereas the rest were considered outliers and removed. The same filtering was applied based on molecular weight, leading to removal of molecules $\leq$200 and $\geq$1000 Da. Table 2 describes the literature assay tasks and correlation between K_p values, and the rest of tasks is reported in Table S2. The aggregated literature data set is available as the Supporting Information.

Table 2. Literature Data Set Description[a]

assay tasks	# CPDs original set	# CPDs after curation	source
Log$P$	12,481	7576	Ulrich et al.[30]
Log$D$	4200	4043	Wu et al.[20]
passive permeability Caco-2 Papp a to b	2582	2475	Wenzel et al.[31]
MDR1 efflux ratio	2298	1956	Esposito et al.[29]
in vivo brain penetration (K_p)	1058	836	Meng et al.[11]

[a] Reported are the assay tasks for the literature MT model, including numerical K_p values, the number of compounds per task (# CPDs), and the corresponding data source.

Figure S1a shows an overview of the internal Novartis and literature data sets based on Uniform Manifold Approximation and Projection (UMAP)[32] with Tanimoto as the similarity metric.[33]

Data Splitting. Model evaluation was carried out with two splitting methods based on temporal information (temporal split) or structural diversity (scaffold split). Temporal information was only available for internal Novartis data, which was split according to compounds' registration date in the internal database. Figure S1b reports a UMAP visualization for Novartis data colored according to temporal information. Compounds registered from 2019 to 2021 constituted the *prospective validation set* for final model evaluation and were always excluded from model training and optimization. Scaffold data splitting was carried out for both internal (except to the prospective validation set) and public data. The Bemis—Murcko definition was utilized to extract molecular scaffolds from the compound data,[34] as illustrated in Figure S2. The internal (and literature) brain penetration data sets were composed by 1332 (477) scaffolds, with an average number of 3.6 (3.4) compounds per scaffold and 1030 (327) singletons. Molecules were split into training and *test sets* with percentages of 90 and 10%, respectively. Five independent scaffold splits were carried out. For early stopping of the GNN training, 10% of the training set was utilized (scaffold split). Using a UMAP visualization, an exemplary scaffold split is shown in Figure S1c for the Novartis compound data.

ML Algorithms. *Univariate Linear Regression.* The relationship between a response and an explanatory variable is modeled by fitting a linear equation of the form $y_i = \beta_0 + \beta_1 x_i$ to the training data, where x_i is the observed value, y_i is the predicted value, and β is the estimated coefficient.[35]

K-Nearest Neighbors. The k-nearest neighbors (k-NN) algorithm classifies a data point using the majority class among the k closest training data points.[36,37]

Random Forest. A random forest (RF) model is an ensemble of multiple decision trees.[38] The variance of individual trees is reduced by bootstrap aggregation. To minimize the correlation between trees, a subset of features is randomly selected at each node. Regression RF prediction was the mean predicted value across trees, whereas majority voting was considered for classification. RF models can handle high-dimensionality data, ignoring irrelevant descriptors, and can be considered as an intrinsic method for feature selection.[39,40] Features' influence in the predictions was estimated using Gini importance or a mean decrease in node impurity.[41,42]

Graph Neural Networks. This method makes use of the molecular graph representation, where nodes represent the atoms and edges correspond to the bonds between the atoms.[15,27] A list of atom descriptors is selected to describe each node, and the neighbor's information can be utilized to apply a set of functional transformations and iteratively merge information from more distant atoms. Here, an open-source implementation of a directed message passing neural network was used.[15] This GNN architecture includes a message passing phase, which is used to build a learned representation of the molecule, and a readout phase, which uses the final representation to make the predictions and is based on a feed-forward neural network with two dense layers constituted of 300 neurons. Three message passing steps were carried out, and a latent representation of dimensionality 300 was extracted. In this implementation, messages are associated with the bonds (edges) instead of the atoms (nodes). The learning rate was linearly increased from 0.0001 to 0.001 for two epochs and then exponentially decreased until 0.0001.[43] Rectified linear unit (ReLU) was utilized as activation function in hidden layers and linear activation for the output nodes. Here, both ST and MT variants of GNNs were generated, which only differed in the number of output units. ST-GNN models had a single neuron in the output layer, whereas MT-GNN models had as many output units as tasks. Ensemble models constituted by five GNNs were utilized, and predicted values were given by the average prediction across the individual GNNs, which only differed in weight initialization.[15]

The scikit-learn[44] library was utilized for univariate linear regression, k-NN, and RF, whereas the Chemprop library was used for GNN models.[15,45] Table S3 reports exemplary scripts to train a model and make predictions.

Hyperparameter Tuning. For RF and k-NN models, we performed a five-fold random cross-validation for hyperparameter tuning on the training set using scikit-learn.[44] For RF regression and classification, the number of trees was set to 100, whereas other hyperparameters were optimized. Candidate values are listed in the following: minimum samples to consider at splitting nodes (2, 5, and 10) or leaf nodes (4, 8, and 12), the maximum depth of the trees (4, 6, 8, and 16), the maximum number of features to consider for node splitting (square root or logarithm in base 2 of the total feature number), and the criterion to evaluate node splitting (mean absolute error or mean squared error for regression, and Gini or entropy for classification). For k-NN classification, the number of neighbors (ranging from 1 to 10) and the weight function used in compound prediction (uniform weights or distance) were tested. Balanced accuracy and coefficient of determination were utilized as performance metrics in the grid search for RF classification and regression, respectively. GNN model optimization was not carried out since the original publication showed only potential marginal improvements with hyperparameter tuning.[15]

Molecular Features. A set of 208 two-dimensional (2D) physicochemical descriptors and Morgan fingerprints (MFPs) with radius 2 and length 2048 were used to represent compound structures as an input for standard ML models.[46] For k-NN, pairs of highly correlated features were removed (Pearson's correlation coefficient > 0.9). The remaining features were centered and scaled to unit variance, also removing those with zero variance. For GNN, the molecules were represented as 2D graphs, and atom and bond features were calculated with RDKit.[15]

Rule-Based Cut-Offs and BBB-Score. *Rule-Based Cut-Offs.* For defining the rule-based cut-offs, the definitions from Xiong et al.[3] were used, that is, <450 molecular weight, 2−4 calculated Log$P$ ($c$Log$P$), 2−3 calculated Log$D$ ($c$Log$D$), <3 H-bond donors, 6−10.5 pK_a, and <90 topological polar surface area (tPSA). Molecular weight, H-bond donors, and tPSA were computed with RDKit,[28] Log$P$/D values were calculated using internal Novartis models, and pK_a was calculated using an internally modified version of MoKa (Molecular Discovery Ltd.).[47−49] Correlation between these physicochemical parameters and experimental K_p values is reported in Table S4 both for internal and literature data.

BBB-Score. The BBB-score was re-implemented in python using RDKit for descriptors, and MoKa for pK_a calculation.[47−49] The model formula was not reparameterized. Molecular weight, the number of aromatic rings, tPSA, and heavy atom count from RDKit did not show significant differences to the Weaver et al.[10] data, whereas H-bond donor and acceptor counts differed between RDKit and the ChemAxon software.[50] A new set of definitions (Table S5) was constructed to minimize the differences in the descriptor values and hence the overall BBB-score. Only 8% of the compounds (91 cases out of 1088) were assigned to a different BBB-score class than in the original paper, where small differences were magnified by rounding (Figure S3). The errors seem to arise from overestimating the number of H-bond donors in the original data, for example, taurosteine (with four H-bond donors reported) or prulifloxacin (eight H-bond donors reported). Errors were also detected in the SMILES for diloxanide, tipranavir, and melitracen. The BBB-score implementation is reported in the Supporting Information (Table S6).

Metrics for Model Evaluation. Regression models' performance was evaluated using the coefficient of determination (R^2), Spearman's correlation coefficient (r_s), and mean absolute error (mae).

$$\text{mae} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value.

The performance of classification models was evaluated using MCC,[21] sensitivity (SE), specificity (SP), and precision in the two classes: positive predictive value (PPV, precision of the positive class BP+) and negative predictive value (NPV, precision of the negative class BP−).

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

where TP is the true positives, TN is the true negatives, FP is the false positives, and FN is the false negatives.

■ RESULTS AND DISCUSSION

ML models for the prediction of brain penetration were developed with internal data from Novartis and public literature data. Model evaluation was carried out with two data splitting settings: scaffold split and temporal split. Scaffold split ensures that the training and test compounds do not share chemical scaffolds and allows evaluating model generalization ability on public data. The temporal data split is a more realistic validation scenario for model usage in the pharmaceutical industry, but it is only applicable to data sets with reported compound registration or measurement date, which was available only for Novartis internal data. Herein, model selection and method benchmark were carried out based on scaffold split both for the internal and public data sets. Next, the model performance was evaluated with a prospective validation set of Novartis internal data (see the Materials and Methods).

Regression Models Based on Internal Data. ML models were trained, and their predictive performance was estimated on five independent test sets using the scaffold split. For such analysis, 2103 compounds from the Novartis internal database with experimental in vivo brain penetration data and a registration date earlier than 2019 were used.

Figure 1 reports the average performance and standard deviation for ML models based on ST and MT learning. ML regression models of different complexities were built for the prediction of LogK_p, including a linear regression using tPSA, a RF regression with physicochemical 2D descriptors or MFP, and a ST-GNN. As a control calculation, we also estimated the predictive performance of predicting the mean LogK_p value of

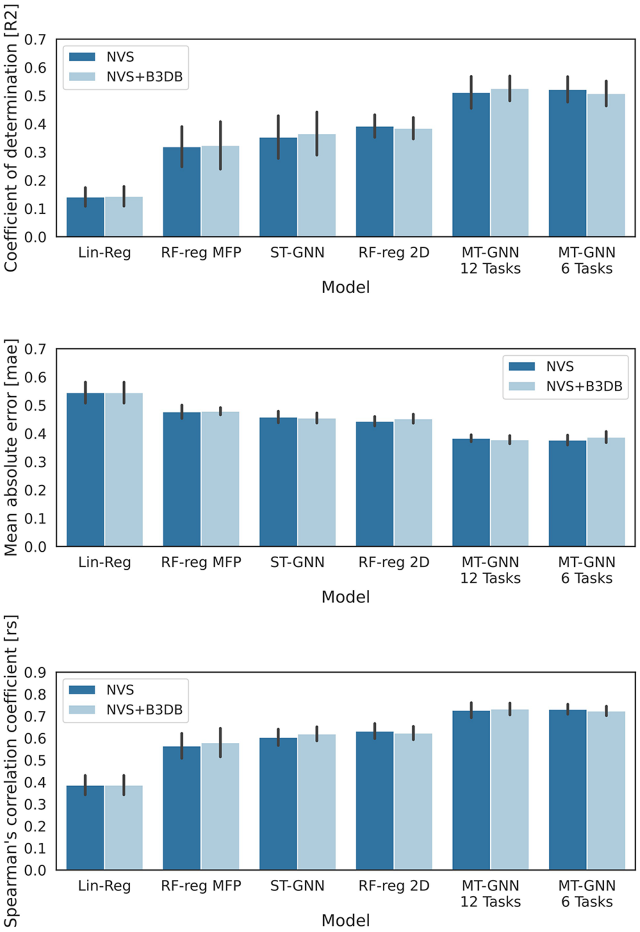


Figure 1. Regression performance of ST and MT models trained with internal data. Reported are the average and standard deviation (error bars) of R^2, mae, and r_s metrics across five independent scaffold splits. Results are shown for six models: RF based on MFP and 2D descriptors (RF-Reg MFP and RF-Reg 2D), linear regression with tPSA (Lin-Reg), ST-GNN, and two MT-GNNs (with 6 or 12 tasks). Models were trained on Novartis brain penetration data (NVS; dark blue) or with the addition of literature data (NVS + B3DB, light blue).

the training set. The performance of all ST models was better than such baseline, which had a mae = 0.61. The RF regression with 2D descriptors outperformed all other ST models, with R^2 = 0.39, mae = 0.44, and r_s = 0.63.

Deep neural networks were utilized for MT learning, that is, to model multiple tasks or properties simultaneously.[25,51] Related tasks to brain penetration were identified and included in a MT model. Since brain penetration predictions is the central goal, we refer to all additional in vitro tasks as "auxiliary tasks". The considered tasks and data size for each of these tasks are reported and described in Table 1. In short, four in vitro passive permeability assays were utilized: PAMPA, LE-MDCK (two assay versions), and Caco-2. Additionally, the MDR1 efflux ratio was considered by including data from the MDCK-MDR1 assay. The main endpoints of these five assays were auxiliary tasks in a MT model with a total of six tasks (*6-task model*). In a second model, six additional auxiliary tasks were included, consisting of secondary endpoints of the above-mentioned assays and lipophilicity (Log*P*, Log*D*) measurements. This second model was constituted by 12 endpoints or tasks (*12-task model*).

As reported in Figure 1, MT learning strategies showed superior performance for the prediction of brain penetration compared to ST. For LogK_p prediction, the best MT model (six-task model) achieved an average performance of R^2 = 0.52, mae = 0.38, and r_s = 0.73 across the five independent scaffold splits. The performance of 6- and 12-task models was equivalent. Thus, the addition of ~110,000 experimental data corresponding to six endpoints did not lead to better predictions. Figure 1 also reports the prediction performance of ML regression models after the addition of 836 training in vivo K_p data points from the literature set (curated B3DB). No benefits were observed by augmenting the internal data set with B3DB data.

Overall, we observed a better performance of MT-GNN compared to ST models in five independent scaffold splits. The top performance was achieved with a MT model based on brain penetration and five auxiliary tasks from in vitro assays related to permeability and MDR1 active efflux. The inclusion of six additional endpoints (12-task model) did not further improve the prediction performance.

Regression Models Based on the Literature Data. Previous modeling brain penetration studies were generally trained and tested on publicly available data sets only. Using the same modeling approaches as with proprietary data, we developed literature-based models to facilitate comparison to other studies and reproducibility. The curated literature set with K_p values for 836 compounds was used for model training and evaluation in five independent scaffold splits.

Figure 2 reports regression performance metrics for ST and MT models based on public data. All literature-based ST models showed better performance than the baseline (i.e., predicting the average LogK_p in the training set), which gave a mae = 0.65. As observed with internal Novartis data, the RF regression using 2D descriptors had better average performance compared to other ST models, with R^2 = 0.42, mae = 0.47, and r_s = 0.69. We built a MT model with the publicly available data. Due to the data availability, it was not possible to retrieve the exact same tasks and number of compounds as in the internal data set. As detailed in Table 2, lipophilicity (Log*P*/*D*), Caco-2 permeability, and MDR1 efflux ratio data sets were extracted from the literature and included as auxiliary tasks for MT modeling. From the literature-based models, MT-GNN was also the best-performing method. Even though the average performance was consistently superior to other models according to multiple metrics (R^2 = 0.44, mae = 0.46, and r_s = 0.72), the observed performance gain by MT-GNN was smaller compared to the observations made with the internal data. For instance, RF regression with 2D descriptors reached a comparable performance with R^2 = 0.42, mae = 0.47, and r_s = 0.69.

Prospective Model Evaluation. In-house compounds registered from 2019 to 2021 constituted the prospective validation set, which included 111 compounds with in vivo brain penetration data. Best ST and MT models were retrained using all data until 2019 (models based on internal data) or all public data. ST-GNN models were also built to evaluate the benefit of MT learning using the same algorithm. Internal and literature-based models were evaluated on the prospective validation set.

Models Trained on the Internal Novartis Data. Table 3 reports the prospective performance of regression models trained on proprietary data. Among ST approaches, RF regression with 2D descriptors provided predictions with the

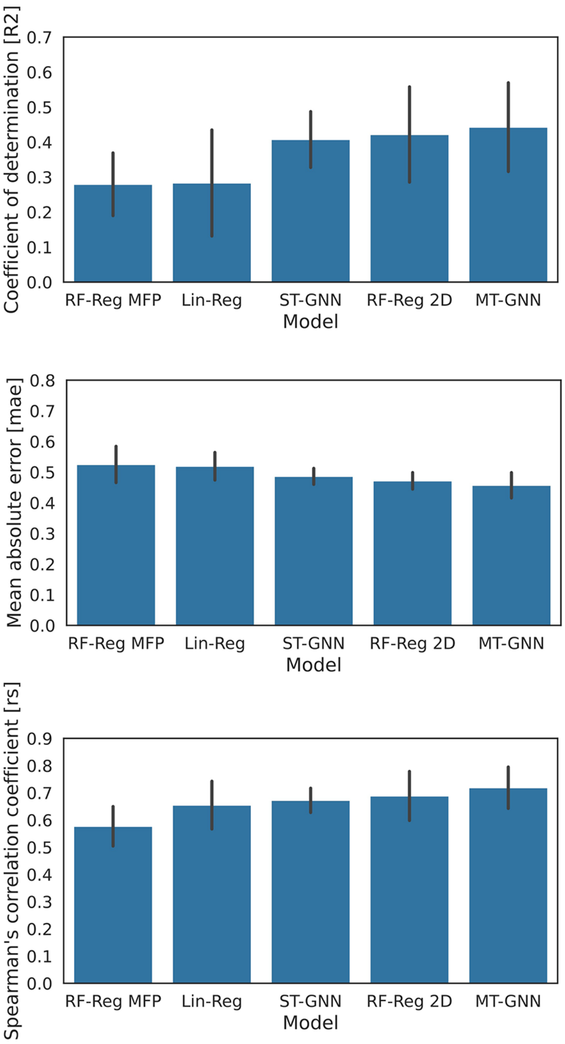


Figure 2. Regression performance of ST and MT models trained with the literature data. Reported are the average and standard deviation (error bars) of R^2, mae, and r_s metrics across five independent scaffold splits. Results are shown for six models: RF based on MFP and 2D descriptors (RF-Reg MFP and RF-Reg 2D), linear regression with tPSA (Lin-Reg), ST-GNN, and two MT-GNNs (with 6 or 12 tasks).

Table 3. Prospective Performance of Regression Models Trained with the Internal Data[a]

	ST RF-Reg 2D	ST-GNN	MT-GNN 6-task model
R^2	0.13	0.10	0.42
r_s	0.51	0.43	0.62
mae	0.48	0.51	0.39

[a] Reported are the R^2, r_s, and mae metrics for the prospective validation set. Results are shown for three models: ST RF regression with 2D descriptors (ST RF-Reg 2D), ST-GNN, and MT-GNN (six-task model with LogK_p brain penetration and five auxiliary tasks reported in Table 1).

lowest error and best compound ranking. Feature relevance in RF models was estimated using Gini importance values, and the most influential descriptors are reported in Figure S4a. MT-GNN (six-task model) achieved the best performance, with R^2 = 0.42, mae = 0.39, and r_s = 0.62. Overall, the prospective model performance was lower compared to the scaffold split setting, especially for ST models.

Models Trained on the Literature Data. Table 4 shows the performance of regression models trained on the literature data when applied to the same prospective validation set. The MT-GNN model based on public data outperformed the ST models. Thus, the benefit of including related auxiliary tasks in a MT-GNN for the prospective prediction of in vivo brain penetration was also observed for models trained with the literature data. Notably, the performance differences between MT and ST learning were more pronounced in this prospective validation set composed of internal Novartis data than in the prediction of compounds with new scaffolds (i.e., scaffold split). These results highlight that depending on the data set and its structural diversity, either a scaffold or a timesplit might be more challenging.[23]

Table 4. Prospective Performance of Regression Models Trained with the Literature Data[a]

	ST RF-Reg 2D	ST-GNN	MT-GNN
R^2	0.19	0.22	0.33
r_s	0.47	0.48	0.44
mae	0.45	0.52	0.58

[a] Reported are the R^2, r_s, and mae metrics for the prospective validation set. Results are shown for three models: ST RF regression with 2D descriptors (ST RF-Reg 2D), ST-GNN, and MT-GNN (literature model with LogK_p brain penetration and five auxiliary tasks reported in Table 2).

Even though ST regression models showed a higher R^2 when trained on public brain penetration data only, our results suggest limited applicability of literature-based ST models for the predictions of internal Novartis compounds. Descriptor relevance was assessed for the RF model, and top-ranked features are reported in Figure S4b. Some prioritized descriptors were common between the RF models trained with the internal and literature data, such as tPSA, hybrid Electrotopological state EState-VSA descriptor 2 (VSA_EState2) and 3 (VSA_EState3),[52] and Log*P*. In both RF models, tPSA was the most important feature. Spearman's correlation coefficient between feature importance values[53] for the internal and literature RF models was 0.83. We observed lower performance of the literature-based MT-GNN model compared to the model relying on internal data, which was further explored by analyzing auxiliary tasks' selection and training data size (vide infra).

Classification Approaches for Decision Making. The experimental variability and modeling error contribute to regression prediction performance. Even though MT-GNNs resulted in consistently improved models compared to ST learning, observed MAE values were 0.38 and 0.39 for scaffold and temporal splits, respectively. To consider these errors during decision-making, accurate categorical predictions might be more useful than continuous values. LogK_p values were categorized into permeable (BP+) and non-permeable (BP−), applying the commonly used LogK_p = −1 threshold (K_p = 0.1). The underlying class distribution was imbalanced, and the prospective validation subset was constituted by 76% BP+ and 24% BP− compounds. To identify successful in silico methods to assist decision-making, we benchmarked classification approaches on the prospective validation set, including ML based on internal data, classical rules with cut-offs, and the BBB-score.

ML Classification. The performance of post hoc classification based on MT-GNN regression was compared to binary classification algorithms. The same auxiliary tasks were used in the regression and classification MT-GNN models corresponding to the six-task model. The MT-GNN$_{reg}$ performance refers to the categorical predictions based on a MT-GNN regression model, whereas MT-GNN$_{class}$ refers to the binary classification model results. For regression models, predicted outputs were transformed to a category by considering LogK_p predictions > −1 (predicted BP+) or ≤ −1 (predicted BP−), as shown in Figure 3. For the MT-GNN$_{class}$ model, median values were used to transform the endpoints of auxiliary tasks into a binary variable.

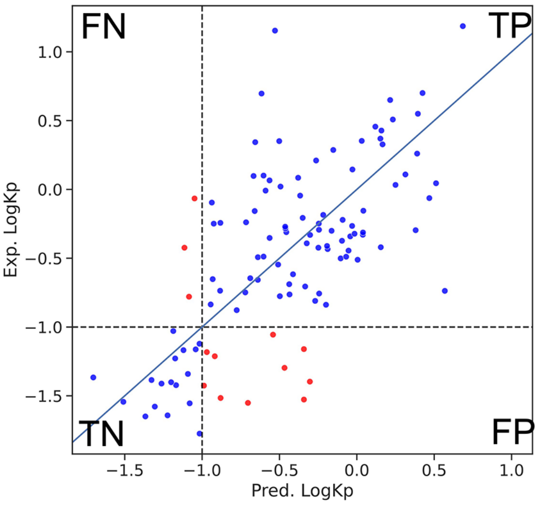


Figure 3. MT-GNN regression-based classification. Predicted (x-axis) vs observed (y-axis) LogK_p values are shown for the prospective validation of the MT-GNN model. Regression predictions are utilized to classify compounds into brain penetrant or non-penetrant with a threshold of LogK_p = −1 (dashed lines). Correct and incorrect classifications are reported in blue and red, respectively.

Table 5 shows the prospective model performance, with MT-GNN regression being the best model, with an MCC = 0.66 and high precision for both classes (PPV = 0.89 and NPV = 0.85). Thus, when the MT-GNN regression model was utilized for brain-penetrant versus non-penetrant classifications, 89 and 85% of brain-penetrant and non-penetrant predictions were correct, respectively.

Interestingly, the MT-GNN regression-based classification outperformed the a priori trained MT-GNN classifier. However, both MT-GNN models were superior to the RF classifier, with an MCC difference higher than 12%. Notably, the predictions for the minority class BP− improved with MT-GNN models, with a NPV improvement over 25% compared to ST models. As a control, a k-NN model was trained and used as a performance baseline. Models were superior to k-NN, suggesting that predictions do not only rely on the category of the most similar training compound (MCC$_{k\text{-NN}}$ = 0.02).

Physicochemical Cut-Offs and BBB-Score. ML models were compared to the performance of more classical medicinal chemistry rules: (i) tPSA cut-off (predicted BP− compound if tPSA > 90 Å^2), (ii) Xiong et al.[3] cut-offs (predicted BP− compound if molecular weight > 450 Da, 2−4 *c*Log*P*, 2−3 *c*Log*D*, <3 H-bond donors, 6−10.5 pK_a, and tPSA > 90 Å^2), and (iii) BBB-score. Gupta et al.[10] described compounds with a BBB-score ≥ 4 as CNS compounds, and compound scoring lower as non-CNS compounds. We applied the same threshold to define compounds as BP+ (BBB-score ≥ 4) and BP− (BBB-score < 4). These approaches showed lower MCC values compared to MT-GNN models. Even though the rule-based cut-off could identify BP+ compounds with a high precision, it was not predictive of the BP− class, as observed by the low NPV value. Interestingly, a cut-off solely based on tPSA could outperform the RF classification on the prospective validation set, and high precision was achieved on the BP+ class. However, low precision for the BP− class was observed with both the tPSA cut-off and the RF model.

Exploring the Benefits of Including In Vitro Data. The impact of including in vitro data for K_p predictions was further analyzed. First, two-task GNN models were built to evaluate the influence of individual assays and training set size on model performance. Different task selections were also compared for literature versus internal data. Last, model stacking was explored as an alternative to MT learning.

Two-Task Models. The two-task GNN models were trained with LogK_p and one auxiliary task at a time (Table S8). Such models improved the ST model results except when using LE-MDCK v2 as the auxiliary task. Besides the model including PAMPA (R^2 = 0.42, mae = 0.38, and r_s = 0.64), all two-task models performed substantially worse than the 6-task model. However, a MT-GNN model with all assays except PAMPA also resulted in a similar performance (R^2 = 0.40, mae = 0.40, and r_s = 0.63), showing the benefits of combining multiple auxiliary tasks and signs of performance saturation. The least and most influential assays were LE-MDCK v2 and PAMPA, which were the smallest and largest assays in terms of data

Table 5. Classification Model Performance on the Prospective Validation Set[a]

	ST RF-Class 2D	ST k-NN 2D	MT-GNN$_{class}$ 6-task[b]	MT-GNN$_{reg}$ 6-task[b]	tPSA cut-off[c]	Rule-Based Cut-Off[d]	BBB-Score[e]
MCC	0.30	0.02	0.44	**0.66**	0.43	0.14	0.18
PPV	0.86	0.76	0.85	0.89	0.91	**1.00**	0.86
NPV	0.40	0.26	0.65	**0.85**	0.47	0.26	0.30
SE	0.68	0.83	0.92	**0.96**	0.71	0.07	0.38
SP	0.67	0.19	0.48	0.63	0.77	**1.0**	0.81

[a] Five classification performance metrics (MCC, PPV, NPV, SE, and SP) are reported for ST models (RF and K-nearest neighbor classification with 2D descriptors), six-task MT-GNN models based on internal auxiliary tasks and data (MT-GNN$_{class}$: classification model, MT-GNN$_{reg}$: regression model with post hoc classification), and three rule-based cut-offs or MPOs (tPSA cut-off, rule-based cut-off, and BBB-score). [b] Passive permeability with PAMPA, LE-MDCK (Papp a to b) two-assay versions, passive permeability Caco-2 efflux ratio, MDCK-MDR1 efflux ratio, and in vivo brain penetration. [c] Predicted BP− compound if tPSA > 90 Å^2. [d] Predicted BP− compound if molecular weight > 450 Da, 2−4 *c* Log*P*, 2−3 *c* Log*D*, <3 H-bond donors, 6−10.5 pK_a, and tPSA > 90 Å^2). [e] BBB-score ≥ 4 classified as BP+ and BBB-score < 4 classified as BP−.

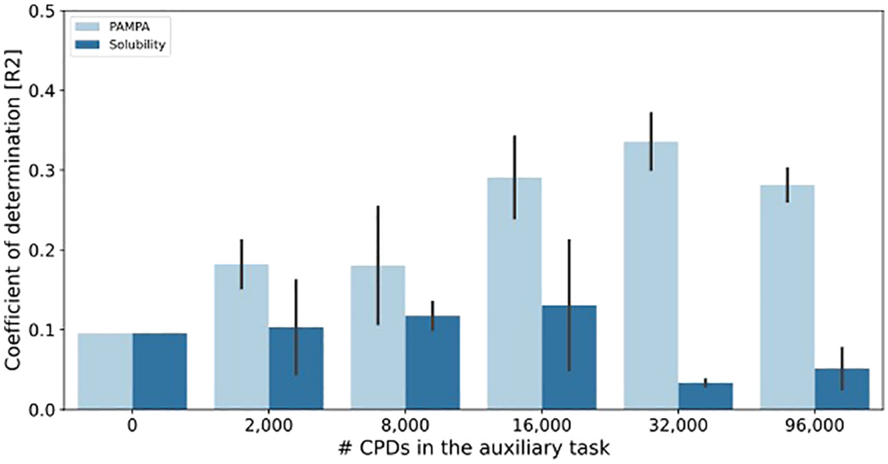


Figure 4. Regression model performance at varying training set size. Prospective R^2 performance of two-task GNN regression models is reported at increasing amounts of auxiliary data. The auxiliary tasks are PAMPA passive permeability (light blue) and solubility at pH = 6.8 (dark blue). The brain penetration main task has a constant training set size of 2103 compounds. Error bars report the standard deviation for three independent trials with different compound subset selection for the auxiliary tasks (from a pool of 117,285 compounds).

quantity, respectively. Therefore, the influence of training set size was further investigated.

Training Set Size. To estimate the effect of data size on MT learning, two-task GNN models were systematically built based on varying amounts of training data. Apart from brain penetration, models included either PAMPA or HT aqueous solubility as an auxiliary task. PAMPA and solubility assays were selected to explore the effect of adding data from a related and an unrelated assay with large data set size. As anticipated, for compounds with overlapping measurements, correlation with experimental LogK_p was larger for PAMPA (r_s = 0.41) than for solubility (r_s = −0.16). Models were built with training sets of increasing size, and three independent trials were carried out. Only compounds with available measurements for both PAMPA and solubility assays were considered for this analysis. Therefore, the two-task models only differed in the auxiliary tasks and not in the compound selection. Figure 4 shows the average R^2 performance for different numbers of training set compounds in the auxiliary tasks. A performance increase was observed with the addition of PAMPA experimental data as the auxiliary task until a saturation effect was observed. The addition of 32,000 training compounds lead to an improvement from $R_0{}^2$: 0.10 to $R_{32k}{}^2$: 0.34, and no significant improvement with the inclusion of more data. This effect was not observed for the model with solubility as the auxiliary task. The addition of solubility data did not yield any performance improvements even with the inclusion of 96,000 compounds.

Literature and Novartis Task Selection. To evaluate whether the observed performance differences between the internal Novartis and literature-based MT-GNN models are arising from the different selection of auxiliary tasks, we built a model with internal data and limited auxiliary data to Log*P*, Log*D*, Caco-2, and MDR1 to match the literature auxiliary task selection. Figure S5 reports the performance of MT-GNN models based on the same auxiliary task selection but trained on internal Novartis or literature data. Even though both models were constituted by the same auxiliary tasks, the training size was larger for the internal Novartis model. The MT-GNN model relying on the internal data was slightly superior to the model generated with the literature data with R^2: 0.37 versus 0.33, mae: 0.41 versus 0.44, and r_s: 0.62 versus 0.58.

Moreover, an experiment was conducted to estimate the effect of replacing the internal brain penetration data by the literature data but keeping the Novartis auxiliary tasks in MT-GNN. Therefore, we built a MT-GNN model with the internal auxiliary tasks of the six-task model (PAMPA, LE-MDCK version 1 and 2, Caco-2, and MDR1) and the public data for the brain penetration task. Figure S5 shows the prospective predictions for two models trained with Novartis auxiliary tasks, being the only difference the LogK_p data employed, that is, Novartis brain penetration or literature data. Both models provided equivalent performance, with R^2: 0.42 versus 0.42, mae: 0.40 versus 0.39, and r_s: 0.64 versus 0.62 for Novartis and B3DB brain penetration data, respectively.

Comparison to Model Stacking. Model stacking was explored as an alternative to MT learning. Five ST-GNNs were built for each auxiliary task (auxiliary models), and their predictions were used as features for a RF that predicted the main task. Two stacked RF models were generated: (i) solely based on auxiliary models' predictions and (ii) with additional 2D descriptors. The prospective performance of RF models is reported in Table S7. The stacked RF model with 2D descriptors showed improved predictions compared to the RF solely based on 2D descriptors, highlighting the information content of the in vitro assay data (R^2 = 0.34, mae = 0.44, and r_s = 0.60). PAMPA and Caco-2 predictions were the most relevant features, whereas tPSA and VSA_EState2[52] were the most influential descriptors. However, stacked models did not reach the performance of MT-GNNs.

Taken together, these results illustrate the performance improvement when including auxiliary in vitro data for compound penetration predictions. Moreover, the importance of auxiliary task and data selection, both in terms of relatedness and data quantity, has shown to be crucial for accurate LogK_p predictions using MT-GNNs.

■ CONCLUSIONS

In this study, we investigated different ML strategies for the prediction of brain penetration with internal and publicly available data. Interestingly, model performance was improved relying on in vitro training data of assays related to brain penetration, as opposed to enriching the training set with public in vivo brain penetration data. MT-GNNs were superior to ST models (models solely based on in vivo brain penetration data) as well as more classical medicinal chemistry approaches, such as cut-offs based on physicochemical properties. Classical medicinal chemistry cut-off rules might be more intuitive for compound optimization or selection. Since they only rely on a small set of common physicochemical parameters such as Log$P$ or molecular weight, interpretability is a clear advantage. Nevertheless, our results have shown that ML leads to more accurate brain penetration predictions. The benefit of including auxiliary tasks in a MT-GNN model was observed both with proprietary and literature data, especially in prospective predictions. However, our results suggested limited applicability of literature-based ST models for the predictions on internal Novartis compounds. Moreover, the superiority of MT learning for brain penetration predictions was observed for numerical LogK_p and categorical predictions across different evaluation scenarios. Accurate predictions on scaffold and temporal splits indicate model robustness and applicability to new chemical series, and performance monitoring is key for a successful model integration in drug discovery.

Taken together, we have shown the benefit of MT-GNN for compound brain penetration predictions, which enable prioritization or reduction of in vivo testing. Precise brain penetration predictions can be utilized for early compound selection, either to achieve brain penetration for CNS drugs or to de-risk compounds that are not intended to act in the CNS. We anticipate MT-GNNs to also be a promising approach to model K_{puu} when more data on this endpoint is available in the future. Our findings indicate that this modeling approach has considerable potential for practical applications in the context of other property predictions. Importantly, our analyses on the effect of task and data selection highlighted the relevance of domain knowledge for proper auxiliary task selection.

■ DATA AND SOFTWARE AVAILABILITY

The data set curated from literature sources, exemplary code to reproduce the MT-GNN model, make predictions, and calculate the BBB-score are made available. Models were also benchmarked and compared based on proprietary data, which are not shared. We hope that the provided data and code will be helpful for improved brain penetration predictions and future work in the field.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00412.

Assay description, correlation of brain penetration data to other assays and physicochemical properties, and data set visualizations; additional results on GNN models' evaluation and feature importance analyses of RF models; code and details on the BBB score open-source implementation (PDF)

Data Set (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Raquel Rodríguez-Pérez − *Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland;* orcid.org/0000-0002-2992-3402; Phone: +41-795-42-2309; Email: raquel.rodriguez_perez@novartis.com

Authors

Seid Hamzic − *Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland*
Richard Lewis − *Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland;* orcid.org/0000-0002-5478-8599
Sandrine Desrayaud − *Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland*
Cihan Soylu − *Novartis Institutes for BioMedical Research Inc., Cambridge, Massachusetts 02139, United States*
Mike Fortunato − *Novartis Institutes for BioMedical Research Inc., Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-1344-5642
Grégori Gerebtzoff − *Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00412

Author Contributions

The study was carried out, and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Vanessa Marx for helpful scientific discussions. This project was part of the Translational Medicine Data Science Academy program at NIBR.

■ ABBREVIATIONS

BBB, blood−brain-barrier; CNS, central nervous system; MPO, multi-parameter optimization; ML, machine learning; BP+, brain penetrant; BP−, non-brain penetrant; MCC, Matthew's correlation coefficient; ST, single-task; MT, multi-task; GNNs, graph neural networks; P-gp, P-glycoprotein; MDCK, Madin−Darby canine kidney; Papp, apparent permeability; k-NN, k-nearest neighbors; RF, random forest; 2D, two-dimensional; tPSA, topological polar surface area; MFP, Morgan fingerprint; R^2, coefficient of determination; r_s, Spearman's correlation coefficient; mae, mean absolute error; PPV, positive predictive value; NPV, negative predictive value; SE, sensitivity; SP, specificity; TP, true positives; TN, true negatives; FP, false positives; FN, false negatives

■ REFERENCES

(1) Abbott, N. J.; Patabendige, A. A. K.; Dolman, D. E. M.; Yusof, S. R.; Begley, D. J. Structure and Function of the Blood-Brain Barrier. *Neurobiol. Dis.* **2010**, *37*, 13−25.

(2) Khan, A. R.; Liu, M.; Khan, M. W.; Zhai, G. Progress in Brain Targeting Drug Delivery System by Nasal Route. *J. Controlled Release* **2017**, *268*, 364−389.

(3) Xiong, B.; Wang, Y.; Chen, Y.; Xing, S.; Liao, Q.; Chen, Y.; Li, Q.; Li, W.; Sun, H. Strategies for Structural Modification of Small Molecules to Improve Blood-Brain Barrier Penetration: A Recent Perspective. *J. Med. Chem.* **2021**, *64*, 13152−13173.

(4) Mensch, J.; Oyarzabal, J.; Mackie, C.; Augustijns, P. In Vivo, In Vitro and In Silico Methods for Small Molecule Transfer Across the BBB. *J. Pharm. Sci.* **2009**, *98*, 4429−4468.

(5) Di, L.; Rong, H.; Feng, B. Demystifying Brain Penetration in Central Nervous System Drug Discovery. *J. Med. Chem.* **2013**, *56*, 2−12.

(6) Bagchi, S.; Chhibber, T.; Lahooti, B.; Verma, A.; Borse, V.; Jayant, R. D. In-Vitro Blood-Brain Barrier Models for Drug Screening and Permeation Studies: An Overview. *Drug Des., Dev. Ther.* **2019**, *13*, 3591−3605.

(7) 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*, 4977−5010.

(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3−26.

(9) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach to Enable Alignment of Druglike Properties. *ACS Chem. Neurosci.* **2010**, *1*, 435−449.

(10) Gupta, M.; Lee, H. J.; Barden, C. J.; Weaver, D. F. The Blood-Brain Barrier (BBB) Score. *J. Med. Chem.* **2019**, *62*, 9824−9836.

(11) Meng, F.; Xi, Y.; Huang, J.; Ayers, P. W. A Curated Diverse Molecular Database of Blood-Brain Barrier Permeability with Chemical Descriptors. *Sci. Data* **2021**, *8*, 289.

(12) Wang, Z.; Yang, H.; Wu, Z.; Wang, T.; Li, W.; Tang, Y.; Liu, G. In Silico Prediction of Blood−Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods. *ChemMedChem* **2018**, *13*, 2189−2201.

(13) Liu, L.; Zhang, L.; Feng, H.; Li, S.; Liu, M.; Zhao, J.; Liu, H. Prediction of the Blood-Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. *Chem. Res. Toxicol.* **2021**, *34*, 1456−1467.

(14) Brito-Sánchez, Y.; Marrero-Ponce, Y.; Barigye, S. J.; Yaber-Goenaga, I.; Morell Pérez, C.; Le-Thi-Thu, H.; Cherkasov, A. Towards Better BBB Passage Prediction Using an Extensive and Curated Data Set. *Mol. Inf.* **2015**, *34*, 308−330.

(15) 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*, 3370−3388.

(16) Plisson, F.; Piggott, A. Predicting Blood−Brain Barrier Permeability of Marine-Derived Kinase Inhibitors Using Ensemble Classifiers Reveals Potential Hits for Neurodegenerative Disorders. *Mar. Drugs* **2019**, *17*, 81.

(17) Alsenan, S.; Al-Turaiki, I.; Hafez, A. A Deep Learning Approach to Predict Blood-Brain Barrier Permeability. *PeerJ Comput. Sci.* **2021**, *7*, No. e515.

(18) Shaker, B.; Yu, M.-S.; Song, J. S.; Ahn, S.; Ryu, J. Y.; Oh, K.-S.; Na, D. LightBBB: Computational Prediction Model of Blood-Brain-Barrier Penetration Based on LightGBM. *Bioinformatics* **2021**, *37*, 1135−1139.

(19) Radchenko, E. V.; Dyabina, A. S.; Palyulin, V. A. Towards Deep Neural Network Models for the Prediction of the Blood-Brain Barrier Permeability for Diverse Organic Compounds. *Molecules* **2020**, *25*, 5901.

(20) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2017**, *9*, 513−530.

(21) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6.

(22) Wu, Z.; Xian, Z.; Ma, W.; Liu, Q.; Huang, X.; Xiong, B.; He, S.; Zhang, W. Artificial Neural Network Approach for Predicting Blood Brain Barrier Permeability Based on a Group Contribution Method. *Comput. Methods Progr. Biomed.* **2021**, *200*, 105943.

(23) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

(24) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4*, 4367−4375.

(25) Montanari, F.; Kuhnke, L.; ter Laak, A.; Clevert, D. A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2020**, *25*, 44.

(26) Rodríguez-Pérez, R.; Bajorath, J. Prediction of Compound Profiling Matrices, Part II: Relative Performance of Multitask Deep Learning and Random Forest Classification on the Basis of Varying Amounts of Training Data. *ACS Omega* **2018**, *3*, 12033−12040.

(27) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2224−2232.

(28) *RDKit v2021.03.5*: Open-Source Cheminformatics. 2006.

(29) Esposito, C.; Wang, S.; Lange, U. E. W.; Oellien, F.; Riniker, S. Combining Machine Learning and Molecular Dynamics to Predict P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2020**, *60*, 4730−4749.

(30) Ulrich, N.; Goss, K. U.; Ebert, A. Exploring the Octanol−Water Partition Coefficient Dataset Using Deep Learning Techniques and Data Augmentation. *Commun. Chem.* **2021**, *4*, 90.

(31) 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*, 1253−1268.

(32) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**, arXiv:1802.03426.

(33) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, *7*, 20.

(34) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(35) Altman, N.; Krzywinski, M. Points of Significance: Simple Linear Regression. *Nat. Methods* **2015**, *12*, 999−1000.

(36) Fix, E.; Hodges, J. L. Discriminatory Analysis. Nonparametric Discrimination: Consitency Properties. *Int. Stat. Rev.* **1951**, *57*, 238−247.

(37) Altman, N. S. An Introduction to Kernel and Nearest Neighbor Nonparametric Regression. *Am. Statistician* **1991**, *46*, 175−185.

(38) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(39) 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*, 1947−1958.

(40) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, 2013.

(41) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Routledge, 2017; Vol. *1−358*.

(42) Ziegel, E. R. *The Elements of Statistical Learning. Springer Series in Statistics*; Springer: New York, New York, 2003; Vol. *45*.

(43) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, *31*, 5999−6009.

(44) Pedregosa, F.; Varoquaux, 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.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(45) *Chemprop: Message Passing Neural Networks for Molecule Property Prediction, (v1.3.1)*, (accessed 2022-04-04).

(46) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(47) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172−2181.

(48) MoKa—pKa modeling. https://www.moldiscovery.com/software/moka/ (accessed March 01, 2022).

(49) Gedeck, P.; Lu, Y.; Skolnik, S.; Rodde, S.; Dollinger, G.; Jia, W.; Berellini, G.; Vianello, R.; Faller, B.; Lombardo, F. Benefit of Retraining PKa Models Studied Using Internally Measured Data. *J. Chem. Inf. Model.* **2015**, *55*, 1449−1459.

(50) ChemAxon. www.chemaxon.com (accessed March 01, 2022).

(51) Rich, C. Multitask Learning. *Mach. Learn.* **1997**, *28*, 41−75.

(52) Kier, L. B.; Hall, L. H. Molecular Structure Description: The Electrotopological State. *Pharm. Res.* **1990**, *07*, 801−807.

(53) Rodríguez-Pérez, R.; Bajorath, J. Feature Importance Correlation from Machine Learning Indicates Functional Relationships between Proteins and Similar Compound Binding Characteristics. *Sci. Rep.* **2021**, *11*, 14245.