

RESEARCH ARTICLE

Open Access



# Blood–brain barrier penetration prediction enhanced by uncertainty estimation

Xiaochu Tong<sup>1,2</sup>, Dingyan Wang<sup>1,2</sup>, Xiaoyu Ding<sup>1,2</sup>, Xiaoqin Tan<sup>1,2</sup>, Qun Ren<sup>3,1</sup>, Geng Chen<sup>1,2,4</sup>, Yu Rong<sup>5</sup>, Tingyang Xu<sup>5</sup>, Junzhou Huang<sup>5</sup>, Hualiang Jiang<sup>1,2</sup>, Mingyue Zheng<sup>1,2\*</sup> and Xutong Li<sup>1,2\*</sup> 

## Abstract

Blood–brain barrier is a pivotal factor to be considered in the process of central nervous system (CNS) drug development, and it is of great significance to rapidly explore the blood–brain barrier permeability (BBBp) of compounds in silico in early drug discovery process. Here, we focus on whether and how uncertainty estimation methods improve in silico BBBp models. We briefly surveyed the current state of in silico BBBp prediction and uncertainty estimation methods of deep learning models, and curated an independent dataset to determine the reliability of the state-of-the-art algorithms. The results exhibit that, despite the comparable performance on BBBp prediction between graph neural networks-based deep learning models and conventional physicochemical-based machine learning models, the GROVER-BBBp model shows greatly improvement when using uncertainty estimations. In particular, the strategy combined Entropy and MC-dropout can increase the accuracy of distinguishing BBB + from BBB – to above 99% by extracting predictions with high confidence level (uncertainty score < 0.1). Case studies on preclinical/clinical drugs for Alzheimer's disease and marketed antitumor drugs that verified by literature proved the application value of uncertainty estimation enhanced BBBp prediction model, that may facilitate the drug discovery in the field of CNS diseases and metastatic brain tumors.

**Keywords:** Blood–brain barrier penetration, BBBp prediction, Uncertainty estimation

## Introduction

With the development of society and the aging of the population, central nervous system (CNS) diseases have become the second largest disease after cardiovascular diseases. However, the success rate of clinical candidate CNS drugs is only about 8%, which is quite low compared with the success rate of 20% for cardiovascular diseases [1]. Cancer is also a major disease in the current society, and the success rate of clinical candidates is only about 5% [2]. Worse still, brain metastases are a common route of disease progression in 20% of patients with cancer.

The vast majority of patients with brain metastases have a poor prognosis even with the treatment of whole-brain radiation therapy. Second-generation kinase inhibitor with blood–brain barrier (BBB) permeability is believed to be one of the effective treatments of brain metastases [3, 4].

Many potential drugs have been discontinued during their development for clinical use for their insufficient quantity to the CNS because the presence of a BBB. BBB is formed by the endothelial cells of the brain capillaries, which controls the transport of molecules between central nervous system and circulatory system, protects the brain from the damage of toxic compounds and maintains the homeostasis inside the CNS [5]. BBB permeability (BBBp) of compounds is affected by many mechanisms. In the clinical applications of CNS drugs, small lipophilic molecules can cross plasmatic membranes to enter the

\*Correspondence: myzheng@simm.ac.cn; lixutong@simm.ac.cn

<sup>1</sup> Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China  
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

brain usually by passive diffusion. Some small hydrophilic compounds are recognized by the endogenous influx or efflux transporters, and some large molecules are undergoing transport through endocytic route [6]. Approximately 98% of small molecules cannot cross the BBB [7]. Inappropriate physicochemical properties (PCP) could limit passive diffusion of drugs into the brain [1], and the active efflux transport, especially the ATP binding cassette family (ABC transporters), could decrease the concentration of many drugs in the CNS by pumping them out from brain to blood. Therefore, it puts forward higher requirements for the design of CNS drugs with not only excellent activity, metabolic properties and low toxicity, but also great BBB penetration that makes them reach the CNS with adequate exposure [8].

With increasing experimental data on BBB permeability, attempts have been made to use computational models to predict the BBBp of compounds, which help to minimize the cost of experiments and facilitate high-throughput screening for enormous compounds. In 1980, Levin proposed the best fit model between the BBB permeability coefficient and logP for compounds with molecular weight (MW) less than 400 Da [9]. Since the early 2000s, there have been a large number of *in silico* BBBp prediction models reported, most of those aimed to find the relationship between scores of molecular physicochemical descriptors and logBB, the unbound brain-to-unbound plasma ratio ( $K_{p,uu}$ ) and  $BBB \pm$  [10–12]. Most of *in silico* predictions have been derived from data on the total brain-to-plasma concentration ratio,  $K_p$ , expressed in its logarithmic form i.e. logBB. The logBB value is affected by the extent of plasma protein and brain tissue binding, however, based on “free-drug hypothesis”,  $K_{p,uu}$  is more informative [6].  $BBB \pm$  is another property used to study the ability of BBBp compounds by dividing compounds into  $BBB+$  and  $BBB-$  groups based on logBB ratio [13, 14] and CNS activity [15], so that the size of the database can be enlarged.

Although many BBBp models have been reported, the number of compounds used to train is very limited, that always leads to overfitting and misleading on the compounds that is outside the chemical space of training data. Many researchers have focused on minimizing the number of molecular descriptors to avoid overfitting, and meanwhile finding physicochemical properties that are crucial for BBBp of compounds to benefit rational drug design. Zhao et al. used 19 simple molecular descriptors for the analysis of 1593  $BBB \pm$  data and showed the importance of hydrogen-bonding properties in modeling BBBp [16]. Gupta et al. built a prediction model “BBB Score”, which consists of stepwise and polynomial piecewise functions. Twenty-two molecular descriptors were studied to describe physicochemical property

space, and five descriptors were selected, namely number of aromatic rings, number of heavy atoms, MWHBN (a descriptor related to MW, hydrogen bond acceptor (HBA) and hydrogen bond donor (HBD)), topological polar surface area (TPSA) and  $pK_a$  [17]. Zhang et al. constructed k-nearest neighbors and support vector machine (SVM) models to predict BBBp using 854 molecular descriptors from different sources, and found that PSA, logP, HBA and HBD are more contributing to the model than others [18]. Yuan et al. showed that the combination of property-based descriptors and molecular fingerprints can significantly improve the performance of SVM-based BBBp prediction model comparing with models using property-based descriptors or molecular fingerprints alone [19]. LightBBB is a BBBp prediction model based on Light Gradient Boosting Machine algorithm with a total of 2432 1D/2D molecular descriptors selected by exclusive feature bundling to avoid overfitting [20]. Alsenan et al. compared kernel PCA, linear PCA, random projection and autoencoder based on BBB dataset with a composed of 6394 property-based descriptors and molecular fingerprints, and proved that dimensionality-reduction techniques can alleviate the overfitting and improve the performance of the model, especially kernel PCA [21]. Roy et al. used 27 molecular descriptors incorporated with 10 molecular solvation energy descriptors based on Kovalenko-Hirata closure (3D-RISM-KH) molecular solvation theory to construct the BBBp prediction model and analyzed the importance of these descriptors by random forest (RF) and gradient boosting machine. A minimum-descriptor-based model with five most important descriptors was obtained, and found it still had good performance [22].

As summarized above, the performance of the above machine learning (ML)-based methods on BBBp prediction depends on the selection of different physicochemical descriptors or molecular fingerprints and subsequent feature extraction, which requires prior knowledge and always prone to bias when selecting features manually. Deep learning (DL) techniques can automatically select optimal features from the provided dataset. In particular, graph neural networks (GNNs) try to learn molecular representation directly from molecular graphs to perform property prediction tasks. Most of these newly proposed GNNs have shown excellent performance based on the evaluation on a benchmark BBBp dataset from MoleculeNet [23]. Xiong et al. proposed Attentive FP for molecular representation by introducing a graph attention mechanism. Attentive FP enabled the graph neural network to extract not only atomic local information but also nonlocal interactions at the intramolecular level to learn additional interactions which affect the overall properties of molecules [24]. Wang et al. built a multichannel

gated recurrent unit architecture to extract molecular features both at the node level and molecule level, so as to cover more elaborate molecular information [25], which may be beneficial to the task of molecular property prediction. Hu et al. established a pre-trained model for unsupervised learning on the graph representation of molecules before using the model to predict molecular properties, which can learn from more unlabeled molecular structures, and to a certain extent, improve the performance of downstream prediction tasks [26]. Recently, Rong et al. proposed a molecular representation framework GROVER [27] based on transformer framework. The strategy of randomly selecting the number of hops can be adapted to different types of datasets, thereby further expanding the learnable molecular datasets. Moreover, node/edge-level and graph-level self-supervised tasks were constructed to learn rich structural and semantic information from a large number of unlabeled molecules in pre-training process. The well-trained GROVER model was fine-tuned with molecules with task-specific labels and used for BBBp prediction as a graph-level task.

For those complex GNNs, simply focusing on the improvement of metrics on the existing benchmark datasets may lead to the neglect of their applicability in practical applications. The deviations between the modeling dataset and the real-world observations will bring the uncertainty for the predictions. For example, in MoleculeNet [23], the molecules defined as BBB+ are about three times as much as the BBB-, while it is estimated that 98% of molecules cannot pass through BBB in the real chemical space [7]. Generally speaking, by adding more high-quality experimental data, the applicability of models can be improved. However, it is not easy to increase the size of BBBp datasets, because measuring the value of BBBp is often complicated, time-consuming and costly. In this circumstance, introducing uncertainty estimation [28] can expand the application domain of the prediction models [29–31]. There are some general approaches for both ML- and DL-based models. For example, Shannon entropy, as a measure of information, also allows us to make accurate statements and perform calculations about the confidence of models' prediction truth [32]. Multiple initialization (Multi-initial) means initializing the model parameters many times randomly to get several independent models trained with same dataset and thus the variances of the prediction results are considered as the uncertainties of the prediction [33]. Recently, there have been some custom-made algorithms of uncertainty estimation to suit the framework of DL model. Monte Carlo dropout (MC-dropout) has been proposed as an approximation of Bayesian neural networks (BNNs) [34, 35] in deep neural networks [36, 37] to reduce computational consumption, which only need

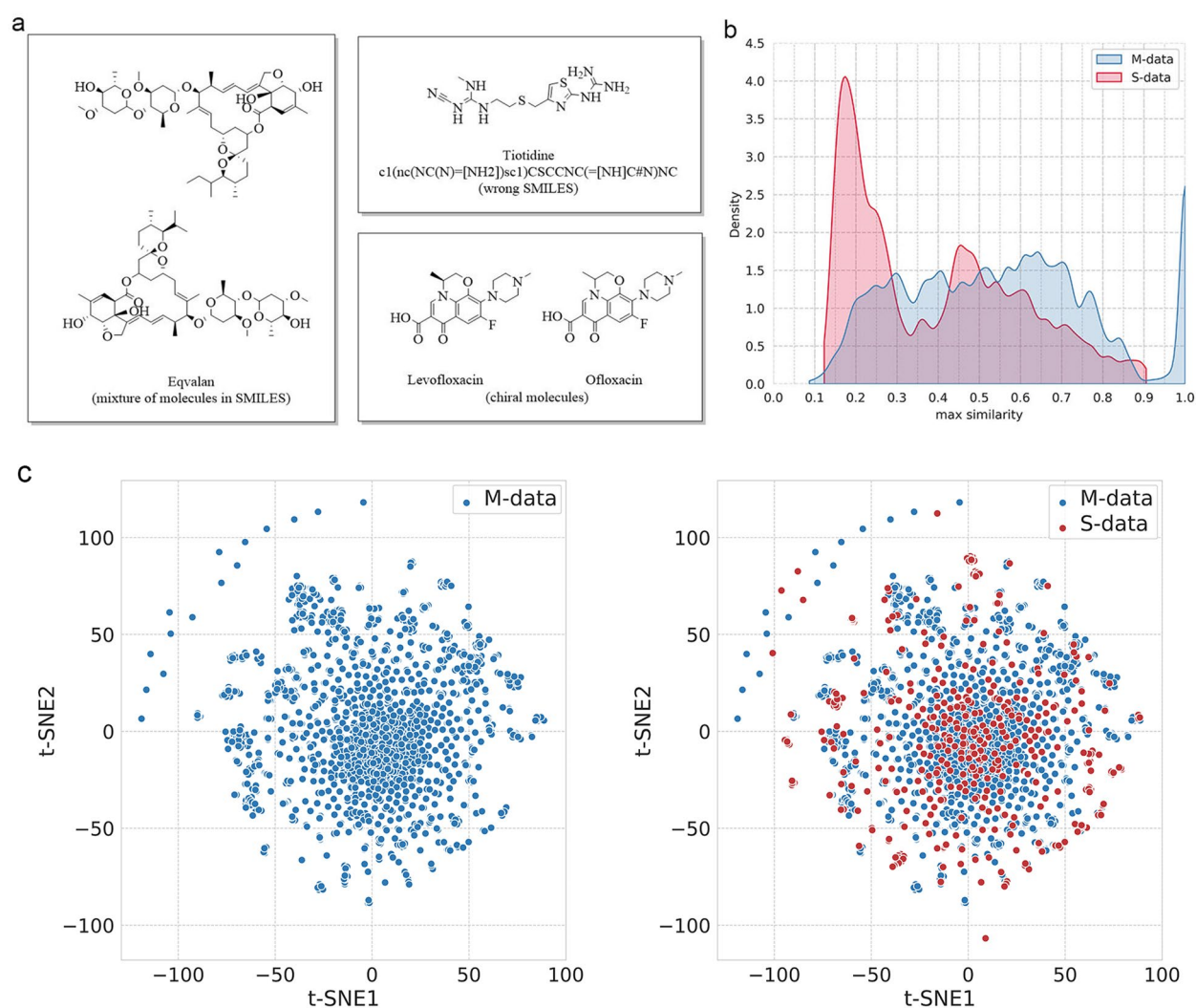
to apply dropout in existing model during inference to get the distribution of prediction results. Besides, compared with the model-agnostic measurement of the distance of molecular representation in a feature space, like fingerprints, the latent space is a more intuitive way to estimate the uncertainty, which does not require retraining the model. Recently, Janet et al. proposed the distance in latent space as a new uncertainty estimation method, which specifically calculates Euclidean distance of each test molecule to the nearest training set molecule in the final layer latent space of deep neural network [38].

In this study, we first analyze the existing BBBp benchmark dataset in MoleculeNet [23], and collect the additional molecules as an external benchmark dataset to evaluate the performance of different types of BBBp prediction models. Furthermore, due to overfitting of DL models and the deviation of training data distribution from the real-world distribution, we introduce uncertainty estimation to quantitatively evaluate the reliability of prediction results and determine the optimal combination of different uncertainty estimation methods. We examine our strategy on preclinical/clinical drugs for Alzheimer's disease and marketed antitumor drugs, and verify it by literature. Ideally, selecting the molecules with certainty for the further wet experiments will reduce unnecessary costs and thus benefit real-world application of *in silico* BBBp prediction model.

## Results and discussion

### Benchmark dataset analysis and new benchmark dataset collection

MoleculeNet [23] is a benchmark for molecular machine learning that curates multiple public datasets focus on different levels of properties of molecules, including a BBBp dataset [39]. The BBBp dataset contains 2053 molecules that were collected from previous works discussing BBB penetration [14, 16, 18, 40]. The molecules are defined as BBB+ or BBB- according to  $\log BB \geq -1$  or  $\log BB < -1$  ( $K_p \geq 0.1$  or  $K_p < 0.1$ ). Although MoleculeNet-BBBp dataset is a relative standard and comprehensive collection of BBBp data, some defects should be pointed out (Fig. 1a). (1) It contains a number of mixtures of molecules like eqvalan. (2) It contains molecules with wrong SMILES that can't be identified by RDKit, like tiotidine. (3) It contains duplicate molecules. For example, ofloxacin is exactly the same molecule named 40,730, and it can also be a duplicate of levofloxacin as some models cannot recognize chiral molecules. Even introducing the information of chirality, enantiomers in the dataset like ofloxacin/levofloxacin can lead to inflated model performance. Thus, we have further processed MoleculeNet-BBBp dataset by removing salts and solvents, neutralizing, and extracting the single molecule with the largest molecular



**Fig. 1** Analyzing molecules' defects in M-data and the distribution of chemical space of S-data and M-data. **a** A list of defective molecules in M-data. **b** The distribution of max similarities inside M-data (blue) and max similarity of each molecule in S-data relative to M-data based on ECFP4 (red). **c** t-SNE distribution of M-data and S-data based on ECFP4

weight from SMILES. After standardization and removal of duplication, we have an updated BBBp benchmark, called M-data, that contains 1937 molecules, comprising 1476 BBB+ and 461 BBB− for further analysis.

After that, we curated an independent supplementary dataset to test the existing BBBp prediction models, named S-data. On the one hand, we supplemented hundreds of CNS and non-CNS drugs that not included in M-data based on Anatomical Therapeutic Chemical (ATC) classification system [15, 41]. ATC annotation is a reasonable inference about whether drugs can cross the human BBB, and can be used to test the generalization performance of computational BBBp model that build on the heterogeneous experiment data of other species.

On the other hand, we collected previous reported compounds from literature [42–44] that are not contained in M-data, and newly released compounds from ChEMBLdb25 [45] that have measured logBB or  $K_p$ . Finally, the new benchmark dataset S-data contains 527 molecules that are assigned to 395 BBB+ and 132 BBB− according to  $\log BB \geq -1$  or  $\log BB < -1$  or CNS/non-CNS drugs. To analyzing the distribution of chemical space of S-data and M-data, we use the Tanimoto similarities and t-distributed stochastic neighbor embedding (t-SNE) based on molecular fingerprint ECFP4 [46]. Figure 1b shows the distribution of max internal similarities in M-data and the distribution of max similarity of each molecule in S-data relative to M-data. The majority of molecules



in S-data are structurally different from M-data with low max similarities range from 0.1 to 0.3. Meanwhile, the visual analysis of t-SNE shows that some molecules in S-data do not overlap with M-data (Fig. 1c). Therefore, S-data can be used as an independent benchmark dataset to measure the generalization ability of BBBp prediction models build on M-data for the deviation in chemical space.

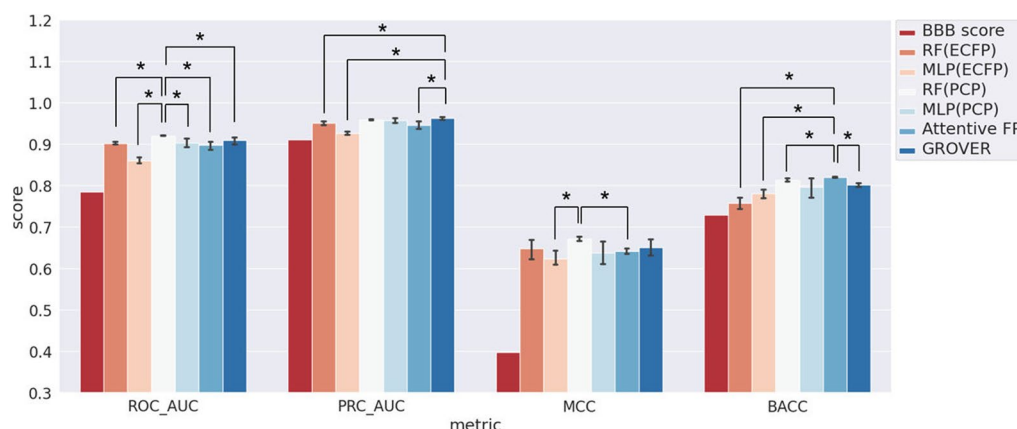
#### Validation of BBBp prediction models on independent S-data

We evaluate the performance of existing *in silico* BBBp prediction models, including BBB score, RF, multi-layer perceptron (MLP), and two the state-of-the-art (SOTA) GNN algorithms, namely Attentive FP and GROVER. First of all, to verify we have correctly rebuilt the GNN-based models, we implement threefold cross validations on these models using M-data as it has done in GROVER. As a self-supervised GNNs model that pre-trained with a large number of drug-like molecules, GROVER shows the best performance with excited metric scores that are in accord with it was reported in its research (the area under ROC curve (ROC\_AUC)=0.976, the area under PRC curve (PRC\_AUC)=0.994, matthews correlation coefficient (MCC)=0.842 and balanced accuracy (BACC)=0.910). Close behind GROVER is Attentive FP, and both of the GNN-based model significantly outperforms RF and MLP (Additional file 1: Table S1).

Next, we focus on the evaluation on S-data. Except BBB score that was a series of predefined linear functions, models were trained by M-data with 5 times runs of different initialization and evaluated by S-data. The implementation details are described in Method, and the performance is shown in Fig. 2 and Additional file 1:

Table S2. Above all, compared to the evaluation results on M-data, the slump in all metric scores of these models substantiates the independence of S-data. Inevitably, BBB score shows much worse generalization ability on the independent S-data, as its selection of molecular descriptors relies on prior knowledge. Among models, GROVER shows highest PRC\_AUC score than others, and significantly higher than RF(ECFP) and MLP(ECFP) model. Attentive FP shows best performance and significantly exceed RF(PCP) when measured by BACC, but RF(PCP) also could be the best according to ROC\_AUC and MCC. Actually, except ECFP-based ML models, RF(PCP), MLP(PCP), Attentive FP and GROVER show moderate and comparable performance on independent testing dataset, i.e. S-data. This rises a suspicion that the exciting metric scores on M-data of these models, especially the SOTA GNN-based models, are only attributed to overfitting.

In fact, it is expected that GNN-based models could learn more task-specific features directly from topological structure of molecules. Thus, we implement a test based on substrates of transporters. Using physicochemical properties as molecular features to predict drug's BBBp is based on an assumption that the majority of drugs could get across the BBB by passive diffusion [47], but active transport mechanisms also exercise considerable influence over the drug concentration in the CNS. The substrates of transporters [48–50] in S-data were extracted to evaluate the performance of these models (Additional file 1: Tables S3 and S4). As a self-supervised-based model that has been pre-trained with a tremendous number of drug-like molecules, we expected that GROVER could distinguish substrates of transporters better than others. However, as the confusion



**Fig. 2** Prediction performance on S-data by BBBp prediction models. Each histogram with an error bar indicates the mean and variance of 5 runs of the model, respectively. Statistical t-tests were applied between the model with the highest metric score and others, and statistically significant test results were noted (\* $p < 0.05$ )

matrix for the prediction results that shown in Table 1, the right predictions (true positive (TP) and true negative (TN)) of PCP-based ML models are on a par with GROVER, as well as Attentive FP. Molecular graph features or pre-training could not give GNN-based models distinct advantages on the prediction of substrates of transporters.

Overall, the moderate improvements of GNN-based models are inconsistent with the hype in their publications. The insufficient modeling data may lead to the inability to give full play to the advantages of DL models in the scenario of BBBp prediction.

#### Commonly used uncertainty estimation in BBBp prediction

In view of the above-mentioned results on independent S-data, the performance of in silico BBBp prediction models is not only limited by algorithms but also by data. Overfitting may undermine the generalization performance of the SOTA DL models with complicated architecture and high-capacity. Furthermore, the deviation between modeling data and real-world data could affect the model's practicability. In fact, it is estimated that about 98% of small molecules are BBB-impenetrable [7], whereas in MoleculeNet-BBBp dataset, BBB+ molecules are over three times as much as BBB- molecules. The practicability of an in silico model that too closely fit to a dataset like that, despite the high accuracy, seems to be rather dubious. In DL models, uncertainty estimation is increasingly important component of assessing prediction truth for its potential to secure its practicability, and would be the high road to circumvent the data obstacle. Here, aiming at bring DL-based BBBp models into play, we focus on how uncertainty impacts the performance of learning from insufficient BBBp data, in particular under the framework of GNNs.

We implemented five proposed algorithms of uncertainty estimation in GROVER-BBBp and Attentive FP-BBBp models. And we also explored uncertainty on PCP-based RF and MLP rather than ML models based on ECFP as the former showed much better performance on BBBp dataset than the latter. Entropy, MC-dropout and Multi-initial can capture the prediction uncertainty of a classification model without any change to its

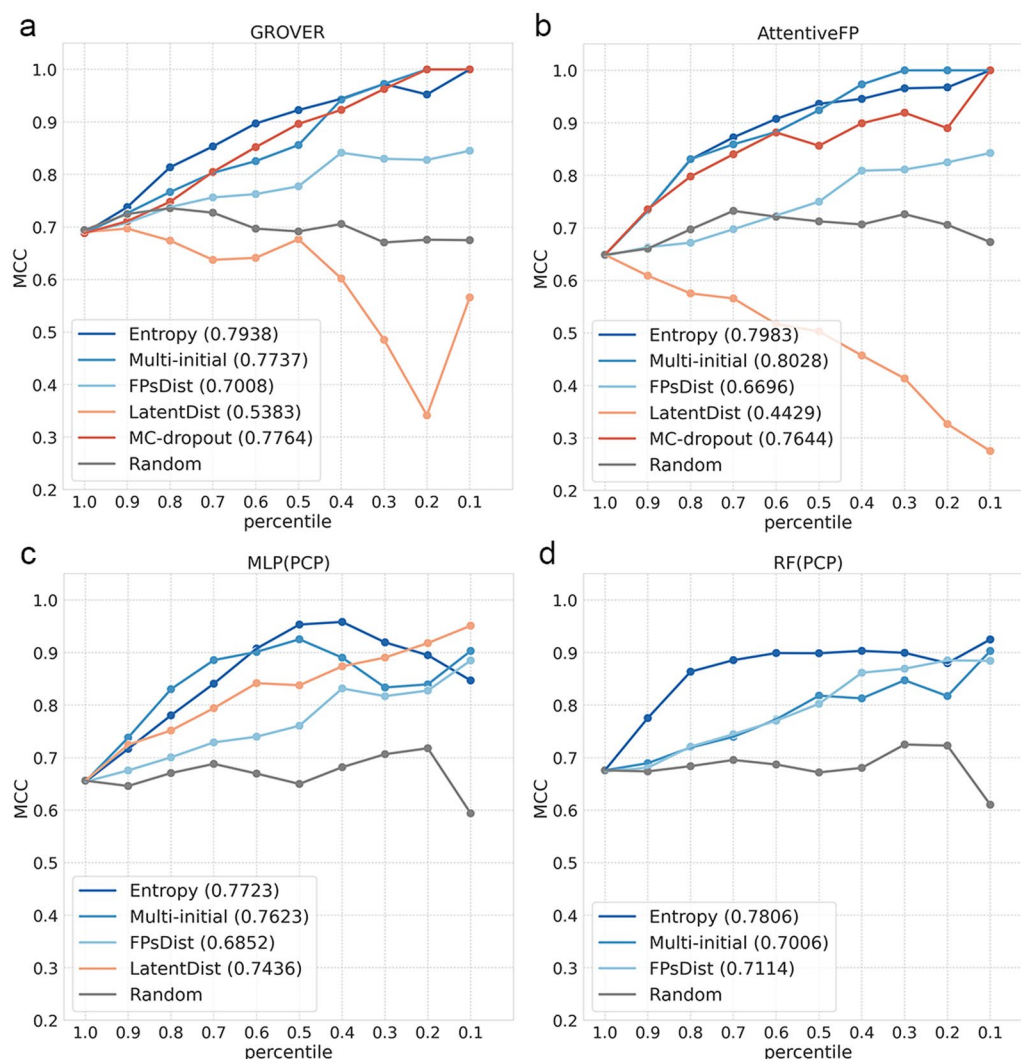
architecture; FPsDist and LatentDist measuring the distance of molecular representation in a feature space and the latent space respectively, are more intuitive ways to estimate the uncertainty as they don't involve re-running or re-training the model. Besides, random method was applied as a baseline for the comparison of different uncertainty estimation methods (See "Methods" Section for more details).

In order to explore the correlation between uncertainty level and prediction correctness, we discarded predictions with top 10% uncertainty in S-data sequentially and calculated the MCC of the remaining (Fig. 3), since MCC is a more stringent metric for imbalance dataset like BBBp. First of all, in all models, the flat trend of the random method indicating that the uncertainty values assigned randomly cannot lead to model improvement, and thus can served as baseline. Entropy, MC-dropout and Multi-initial show relatively better performance with the higher under curve area of MCC (MCC\_AUC) than distance-based methods i.e. FPsDist and Latent-Dist. In particular, for GNN-based model i.e. GROVER-BBBp and Attentive FP-BBBp, Entropy, MC-dropout and Multi-initial lead to relatively steady improvements in MCC with decreases in quantity of high-uncertainty compounds, whereas the MCC curves for MLP(PCP) and RF(PCP) cannot keep rising with these methods. By comparison, FPsDist can supplement relatively robust fingerprint information, and thus shows moderate upward swings in all of models. The trends of Latent-Dist on GNN-based models are the opposite of that on MLP(PCP), probably because of the more serious overfitting to M-data of GNNs than MLP. The closer distance of latent embedding between the training and testing data could exacerbate overfitting and thus cause misleading. Overall, compared to PCP-based ML models, introducing uncertainty estimation to GNN-BBBp models improves performance more greatly, in which the prediction performance measured by MCC can reach 1 for the remaining 10% of the most certain molecules. Furthermore, we used a variety of drug-like datasets to demonstrate the reliability of uncertainty-enhanced GROVER model, and found that Entropy, MC-dropout and Multi-initial can enhance the prediction performance of GROVER robustly in applied 9 binarized datasets from admetSAR [51] (Additional file 1: Tables S8–S19 and Fig. S1).

Considering that GROVER-BBBp model showed relatively good performance when enhanced by varied uncertainty estimation methods containing DL-specific LatentDist and MC-dropout, we focus on GROVER to analyze how the different uncertainty estimation methods and their combinations effect the performance of BBBp prediction task. Figure 4 shows the percentages of molecules

**Table 1** Confusion matrix of model predictions on 27 substrates in S-data

	RF(PCP)	MLP(PCP)	Attentive FP	GROVER
TP	10	11	11	10
FN	3	2	2	3
FP	1	2	3	1
TN	13	12	11	13



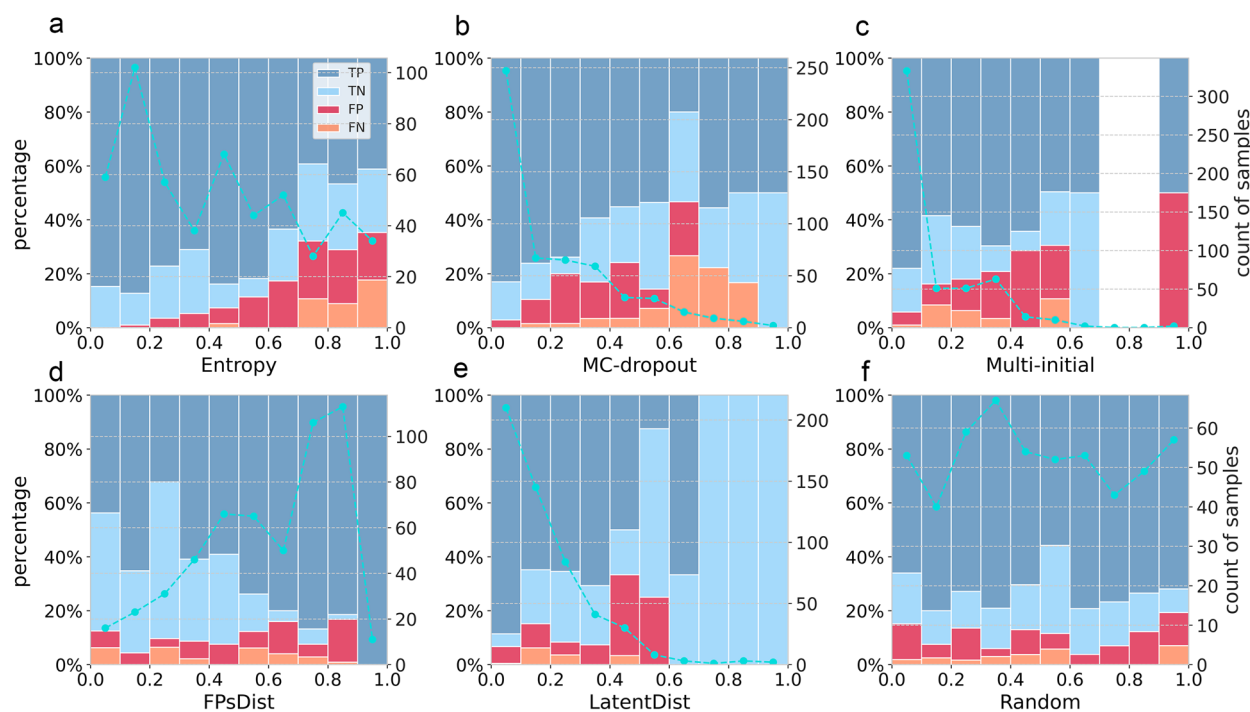
**Fig. 3** Prediction performance by introducing different uncertainty estimation methods for BBBp prediction models. **a** The MCC curves for different uncertainty estimation methods in GROVER, namely Entropy, MC-dropout, Multi-initial, FPsDist, LatentDist and random method. The x-axis is the proportion of remaining compounds in S-data when the compounds with high uncertainty are sequentially discarded, and y-axis is corresponding MCC of the BBBp prediction model. The MCC\_AUC is shown in parentheses. **b** The MCC curves for different uncertainty estimation methods in Attentive FP. **c** The MCC curves for different uncertainty estimation methods in MLP(PCP). **d** The MCC curves for different uncertainty estimation methods in RF(PCP)

in S-data with the prediction of TP, TN, false positive (FP) and false negative (FN) within different uncertainty ranges, and also their total numbers. As shown in Fig. 4a, entropy values of these molecules are comparatively well-distributed in each interval, and as entropy increasing, the proportion of molecules that are incorrectly predicted (FP and FN) gradually increases. When comes to MC-dropout and Multi-initial (Fig. 4b and c), though they tend to give low uncertainty to most of the predicted molecules that makes the distributions of them are non-uniform and even discontinued, the upward trends of the proportion

of misprediction can be also seen when uncertainty get higher. Whereas, the clear trend is not shown in FPsDist and LatentDist, and the latter is even inferior to random method. These observations are in accord with the growth curves of the performance of GROVER that enhanced by these uncertainty estimation methods in Fig. 3a.

#### Optimal combination strategy of uncertainty estimation in BBBp prediction

Next, we attempt to explore whether the ensemble of these uncertainty estimation methods would provide



**Fig. 4** Prediction results from GROVER-BBBp model on S-data within different uncertainty ranges, and corresponding numbers of molecules. **a** Entropy method. **b** MC-dropout method. **c** Multi-initial method. **d** FPsDist method. **e** LatentDist method. **f** Random method

**Table 2** Model performance of various combinations of uncertainty estimation methods in GROVER-BBBp model

Entropy	MC-dropout	Multi-initial	FPsDist	LatentDist	MCC_AUC
✓					0.7938
	✓				0.7764
		✓			0.7737
			✓		0.7008
				✓	0.5383
✓	✓				<b>0.7965</b>
✓	✓	✓			0.7879
✓	✓		✓		0.7956
✓	✓			✓	0.7771

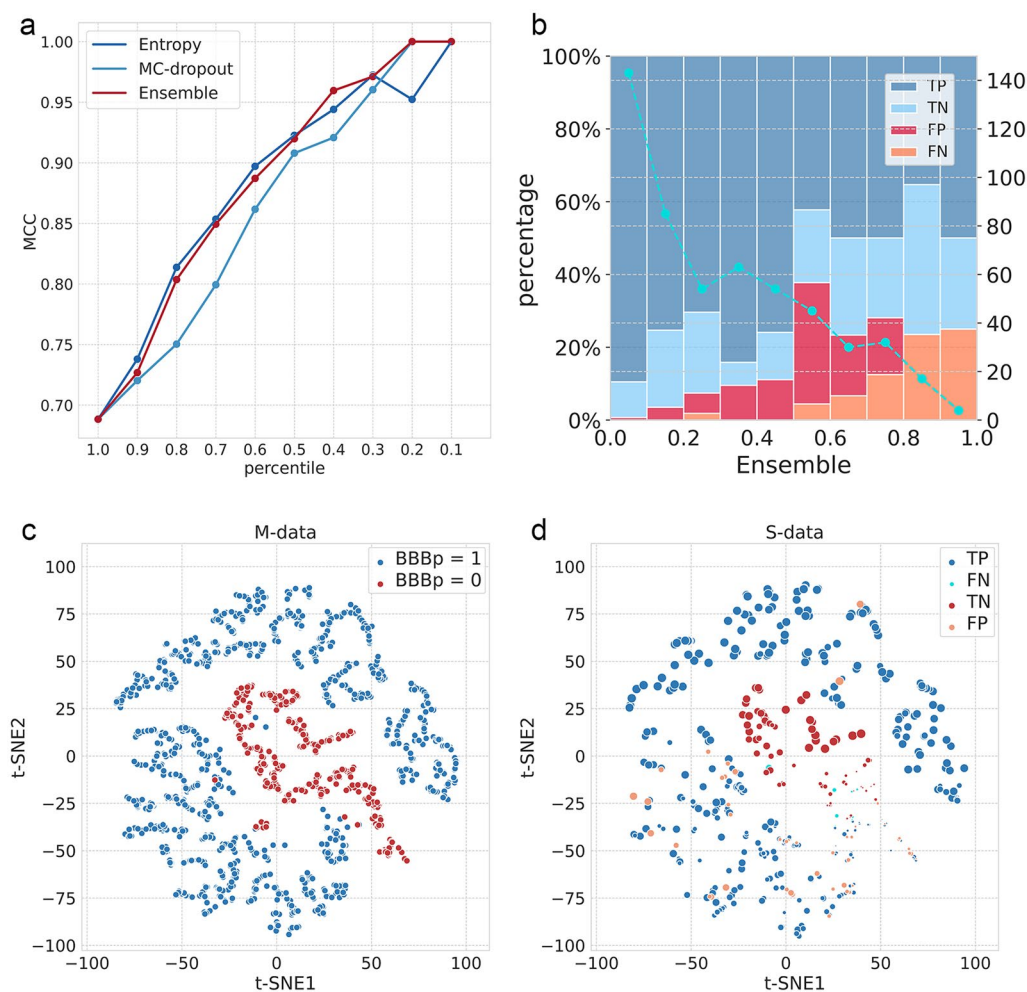
The highest value is highlighted in bold

more steadily enhancement in GROVER-BBBp model. The combined uncertainty values were calculated by moderated-Z (modZ) weighted average algorithm, and the corresponding MCC\_AUC values are summarized in Additional file 1: Table S5 and partly shown in Table 2. It is found that the combination of Entropy and MC-dropout obtains a highest MCC\_AUC among all combination scheme, and outperforms using Entropy method alone, which is the single method of highest MCC\_AUC for GROVER. But MCC\_AUC cannot keep

climbing when add Multi-initial, FPsDist or LatentDist further (Table 2).

Figure 5a shows the MCC curves of Entropy method, MC-dropout method and the ensemble of them. We conclude that the ensemble of Entropy and MC-dropout is a robust strategy for enhancing the model prediction, that has also been verified on other 9 drug-like datasets (Additional file 1: Table S8–S19 and Fig. S1). In particular, when remaining 20% of molecules with lower uncertainty, using the ensemble method shows better performance than using Entropy alone, as the latter shows a slightly decrease here. This improvement is necessary for the reason that molecules with the most certain prediction have priorities over others for further experimental verification. Figure 5b shows the prediction results under the different range of the ensemble uncertainty. For ensemble uncertainty below 0.5 (left side), only 5.4% molecules (22 of 408 molecules) are wrongly predicted. Remarkably, the prediction accuracy is above 99% when ensemble uncertainty values are less than 0.1, as there is only one FP prediction in 144 molecules in range 0–0.1. Therefore, the uncertainty value is a reliable guide for the usage of the GROVER-BBBp model. Predicted BBB+ molecules with lower uncertainty values can be put into experimental verification with greater confidence but





**Fig. 5** Prediction performance by introducing ensemble uncertainty and t-SNE distribution for molecules in M-data and S-data. **a** The MCC curves of Entropy, MC-dropout and ensemble of them. **b** Prediction results of molecules in S-data within different range of the ensemble uncertainty, and corresponding numbers of molecules. **c** t-SNE distribution of M-data based on latent representation of GROVER. **d** t-SNE distribution of S-data based on latent representation of GROVER, and the size of the point represents the uncertainty of the prediction. The larger the size of the point, the smaller the uncertainty value

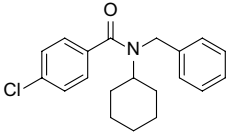
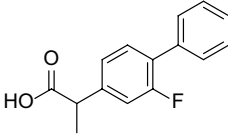
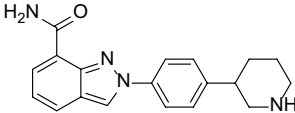
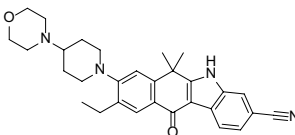
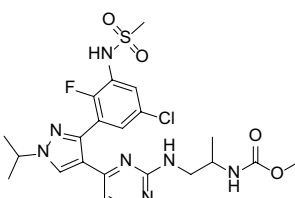
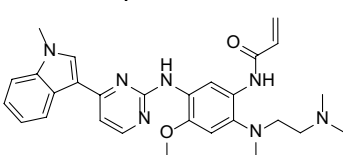
molecules with uncertainty values above 0.5 should be treated with circumspection.

In addition, t-SNE is used to visualize how uncertainty estimation works in GROVER-BBBp model. Figure 5c, d show the t-SNE plots of latent representation of GROVER for molecules in M-data and S-data, respectively. By introducing uncertainty estimation in S-data prediction, the molecules with high uncertainty are mainly concentrated at the junction of BBB+ and BBB-. Thus, uncertainty estimation can help GROVER provide more reliable and practical decisions to further distinguish inseparable molecules in the hidden space.

#### Application of BBBp prediction model enhanced by uncertainty estimation

To verify the practicability of GROVER-BBBp model enhanced by uncertainty estimation, we first predicted preclinical/clinical drugs for Alzheimer's disease, and the results have been summarized in Table 3. FPS-ZM1, a specific RAGE inhibitor to block A $\beta$  binding to the V domain of RAGE, could readily cross the BBB and considered as a candidate drug for the treatment of Alzheimer's disease [52, 53]. Consistently, FPS-ZM1 is predicted to be able to cross the BBB in our model, and the uncertainty value is less than 0.5. Tarenflurbil is predicted to be able to cross the BBB with uncertainty of 0.6328, but its Phase

**Table 3** A list of prediction results with uncertainty of clinical drugs and marketed antitumor drugs

Drug	Structure*	Predicted probability	Uncertainty	Potential indications
FPS-ZM1		0.9761	0.1964	Alzheimer's disease
Tarenflurbil		0.7881	0.6328	Alzheimer's disease
Niraparib		0.9563	0.3676	Carcinoma ovarian, fallopian tube cancer and peritoneal carcinoma
Alectinib		0.9484	0.3907	Non-small cell lung cancer (NSCLC) metastatic, ALK-positive
Encorafenib		0.2529	0.7017	Melanoma with BRAF mutation, colorectal cancer with BRAF V600 mutation
Osimertinib		0.4864	0.7486	NSCLC advanced, metastatic, EGFR mutation

\*Structures of drugs used in model are stripped of chirality

III clinical trial failed due to its poor brain delivery efficiency [54]. In fact, in situ rat brain perfusion experiments have shown that tarenflurbil can rapidly cross the BBB in the absence of plasma protein, but plasma protein binding significantly limits the free plasma fraction of tarenflurbil, and further leads to decrease in the concentration into the brain [55, 56]. As the model was built based on logBB, what only considered the total concentration rather than unbound concentration of drugs in brain and plasma, the high plasma protein binding rate may be the reason for the wrong prediction of tarenflurbil. On the other hand, although mis-predicted, tarenflurbil shows an uncertainty value greater than 0.5, which indicates the low reliability of the prediction result and avoids our excessive trust in it.

Moreover, despite the advances in the treatment of many cancers, CNS tumors and brain metastases still

pose significant challenges, partially because few of the antitumor drugs can penetrate the BBB to reach specific targets in brain. As a test of GROVER-BBBp model, we performed BBBp prediction for some small molecule inhibitors (SMIs) that have been marketed for tumor treatment, and verified by literature. The prediction results of some molecules with available records related to BBBp are shown in Table 3 (the full list of antitumor SMIs and corresponding predictions in Additional file 1: Table S6). In general, because these antitumor SMIs are structurally different from training molecules, their uncertainty values are higher than those in S-data, which are all greater than 0.3. Among these drugs, niraparib and alectinib are predicted as BBB+ with high confidence with uncertainty of 0.3676 and 0.3907, respectively. Niraparib is an FDA-approved poly (ADP-ribose) polymerase-1/-2 inhibitor for anticancer treatment [57, 58].

Niraparib (probability of 0.9563 with uncertainty of 0.3676) has shown its ability in a BRCA2-mutant intracranial tumor model [59], and its brain to plasma exposure ratio is about 0.3, that corroborates the confident BBB+ prediction [57]. Alectinib (probability of 0.9484 with uncertainty of 0.3907) has been approved by FDA to treating ALK-positive NSCLC patients for both systemic and intracranial disease [60–62]. Preclinical studies have also proved that alectinib has a high brain-to-plasma ratio and drug permeability in vitro [63]. Encorafenib (probability of 0.2529 with uncertainty of 0.7017) is an example of TN infer. It is a selective BRAF inhibitor which has been approved for treating melanoma, with a brain-to-plasma ratio of approximately 0.004 in mice model [64, 65].

The GROVER-BBBp model enhanced by uncertainty estimation can correctly predict whether molecules can cross the BBB for most clinical or marketed drugs, and further quantitatively indicate the reliability of the model's prediction results. However, it still has some limitations. Although logBB is a widely used parameter to quantify the brain permeability of drugs, the  $K_{p,uu}$  is a more ideally endpoint than  $K_p$ , as it removes the interference of plasma protein and brain tissue binding that affect  $K_p$ . However, publicly available  $K_{p,uu}$  data is not enough for de novo building of in silico BBBp model. Therefore, we have used Friden's  $K_{p,uu}$  dataset [44] to fine-tune GROVER-BBBp model to predict  $K_{p,uu}$ , expecting to fix the model's bias toward  $K_p$  to some extent. Dataset collected from Colclough [3] and Kim's [66] works was constructed to test the fine-tuned model externally, in which molecules defined as BBB+ or BBB− according to  $K_{p,uu} \geq 0.1$  or  $K_{p,uu} < 0.1$ . The results show that the fine-tuned model can correct some misprediction (Additional file 1: Table S7). For example, osimertinib is EGFR inhibitor for the treatment of advanced NSCLC patients with EGFR-mutated and demonstrated efficacy against stable or asymptomatic CNS metastases [67]. Recent experiment has shown that  $K_{p,uu}$  of osimertinib in-vivo rat is 0.21 [3] that characterizes its permeability [68]. In the original prediction results, GROVER-BBBp model incorrectly predicted it as BBB− (probability of 0.4864 with uncertainty of 0.7486), while the latest fine-tuned model successfully predicted it as BBB+, with the corresponding prediction probability of 0.6952. We can expect future exploration on  $K_{p,uu}$  may provide a more comprehensive prospect for the prediction of BBBp.

## Conclusions

In this study, a newly collected independent dataset S-data was used as a benchmark dataset to evaluate the performance of BBBp prediction models, which contained a total of 527 molecules with 395 BBB+ and

132 BBB− respectively. Among various BBBp models, conventional PCP-based ML models and GNN-based models exhibit moderate and comparable prediction performance, which suggests that overfitting to a bias training dataset could undermine the generalization ability of the SOTA GNN-based models with complicated architecture. Thus, in order to secure the practicability of in silico BBBp prediction models, we introduced uncertainty estimation to quantify the reliability of prediction results of these models. GNN-BBBp models enhanced by uncertainty estimation show greater improvement than PCP-based ML models, and we find that using the combination of Entropy and MC-dropout for uncertainty estimation in GROVER-BBBp model is the optimal strategy. Based on this strategy, the BBBp potential of many drugs of Alzheimer's disease and cancer were successfully predicted with a quantitative estimation of prediction reliability and verified by literature.

In conclusion, our study makes the first attempt to offer insights into prediction uncertainty into ML- and DL-based BBBp prediction model. Uncertainty estimation helps determine how much we can trust a prediction result, enhanced by that, the proposed BBBp in silico model can speed up the high-throughput screening and lead optimization of BBBp molecules, and beneficial to the discovery of drugs for the treatment of CNS diseases and malignant tumors with brain metastasis.

## Methods

### Data sets

The BBBp dataset collected from MoleculeNet [23] contains a total of 2053 BBB+/BBB− molecules defining based on whether  $\log BB \geq -1$  ( $K_p \geq 0.1$ ). After removing salts and solvents, neutralizing, and extracting the single molecule with the largest molecular weight from SMILES, the preprocessed and deduplicated M-data contained 1937 molecules, including 1476 BBB+ and 461 BBB−. The new benchmark dataset was derived from CNS and non-CNS drugs based on ATC [15, 41], and compounds with measured logBB in previous literature [42–44] or ChEMBLdb25 [45], and preprocessed in the same way as M-data to get S-data with 527 molecules (395 BBB+ and 132 BBB−). The  $K_{p,uu}$  dataset from Friden [44] was divided into BBB+/BBB− according to whether  $K_{p,uu} \geq 0.1$ , comprising 24 BBB+ and 17 BBB−. And external test set from Colclough [3] and Kim's [66] works contained 18 molecules (4 BBB+ and 14 BBB−).

### Models for BBBp prediction

In this study, we used different types of BBBp prediction models, including BBB Score [17], conventional ML models RF and MLP, and GNN-based models Attentive FP [24] and GROVER [27].

For BBB score, physicochemical descriptors like MW, HBA, HBD, TPSA and number of aromatic rings were calculated by RDKit toolkit and  $pK_a$  was obtained with Epik module of Maestro. The final BBB score was summed according to the functions associated with the above descriptors, and molecules were considered as BBB+ when the scores  $\geq 4$ , otherwise labeled as BBB-.

ML models often use physical- or chemical-property descriptors or molecular fingerprints as the inputs. In this study, we used simple networks RF and MLP, and molecular fingerprint ECFP4 or physicochemical properties used in GROVER were calculated as the representation input to the ML models.

For RF models, three hyperparameters were considered including 'n\_estimators', 'max\_depth', and 'criterion'. The best group of these hyperparameters for RF(ECFP) is set 'n\_estimators' to 100, 'max\_depth' to 50, and 'criterion' to 'entropy'. The best group of these hyperparameters for RF(PCP) is set 'n\_estimators' to 250, 'max\_depth' to 20, and 'criterion' to 'gini'.

For MLP models, four hyperparameters 'hidden\_layer\_sizes', 'max\_iter', 'batch\_size' and 'learning\_rate\_init' were considered. In MLP(ECFP) model, the best group of these hyperparameters is set 'hidden\_layer\_sizes' to [1000, 500], 'max\_iter' to 3000, 'batch\_size' to 64 and 'learning\_rate\_init' to 0.0001. In MLP(PCP) model, the best group of these hyperparameters is set 'hidden\_layer\_sizes' to [1500, 1000, 500], 'max\_iter' to 1000, 'batch\_size' to 16 and 'learning\_rate\_init' to 0.0001.

For retraining Attentive FP model, we followed the optimal parameters given in article [24]. And on the basis of pre-trained GROVER model [27], we further fine-tuned the downstream BBBp prediction model on M-data following the determined hyper-parameters from the article.

### Uncertainty estimation methods

In this study, we implemented five uncertainty estimation methods, including Entropy, MC-dropout, Multi-initial, FPsDist and LatentDist. At the same time, the random method was used as the benchmark to compare different uncertainty estimation methods.

The random method is achieved by randomly assigning value between 0 and 1 to each molecule in S-data as an uncertainty value.

Schwill proposed that entropy can be used as a measure of uncertainty [32]. And it is the most classical way to measure the uncertainty of classification models. Specifically, the definition of Shannon entropy used in this study is as follows:

$$\mu = - \sum_c p_c \log p_c \quad (1)$$

where  $\mu$  is entropy measure, and  $p_c$  corresponds to the probability value of each class in model's output, which is multiplied by corresponding logarithmic value and finally summed. In BBBp prediction model, we used the probability value of the model's output to get the entropy value, that is, the uncertainty, and the larger this value, the greater the uncertainty of the model.

MC-dropout in deep neural networks has been proved to be used as an approximation of BNN [36, 37]. In practice, there is no need to change the framework of existing DL models, but only need to apply dropout during inference. Finally, the variances of different prediction results obtained by multiple inferences are taken as the values of uncertainty estimation.

Multi-initial is a basic method for uncertainty estimation. The model is trained several times independently with different initialization, and the variances of its results can also be considered as uncertainties of the prediction.

For FPsDist, we calculated the Tanimoto distance of the molecules in S-data relative to the nearest-neighbor molecule in M-data using ECFP4 [69]. LatentDist is a new uncertainty estimation method by measuring the distance in latent space, which specifically calculates Euclidean distance of each test molecule to the nearest training set molecule in the final layer latent space of deep neural network [38].

For the combined uncertainty values obtained by different uncertainty estimation methods, modZ weighted average algorithm is used. Specifically, considering entropy uncertainty as a standard value, Spearman's rank correlation coefficients between other uncertainty values and entropy values are used as weights of uncertainty obtained by other uncertainty estimation methods, and uncertainty values with different weights are averaged to obtain the final combined uncertainty values.

### Evaluation metrics

For assessment of BBBp prediction model performance, several metrics are evaluated. The ROC curve takes false positive rate as the x-axis and true positive rate (recall) as the y-axis, and the area under ROC curve is called ROC\_AUC for short in this study. Similarly, the PRC curve uses recall as the x-axis and precision as the y-axis, and the area under PRC curve is abbreviated as PRC\_AUC in this study. The larger ROC\_AUC value and PRC\_AUC value, the better the performance of the BBBp prediction model. The false positive rate, recall, and precision are defined as follows:

$$\text{false positive rate} = \frac{FP}{TN + FP} \quad (2)$$



$$\text{recall/sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (4)$$

Specifically, BBBp dataset has the problem of data imbalance, BACC and MCC are used in this study to evaluate classification model performance, which are relatively balanced metrics considering TP, TN, FP and FN simultaneously. These two metrics are defined as:

$$\text{BACC} = \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) / 2 \quad (5)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

#### Abbreviations

ALK: Anaplastic lymphoma kinase; ATC: Anatomical Therapeutic Chemical; BACC: Balanced accuracy; BBB: Blood–brain barrier; BBBp: Blood–brain barrier permeability; BNNs: Bayesian neural networks; CNS: Central nervous system; DL: Deep learning; EGFR: Epidermal growth factor receptor; FN: False negative; FP: False positive; GNNs: Graph neural networks; HBA: Hydrogen bond acceptor; HBD: Hydrogen bond donor; MCC: Matthews correlation coefficient; ML: Machine learning; MLP: Multi-layer perceptron; MW: Molecular weight; NSCLC: Non-small cell lung cancer; RAGE: Receptor of advanced glycation end products; ROTB: Rotatable bond; RF: Random forest; SMLs: Small molecule inhibitors; SVM: Support vector machine; TN: True negative; TP: True positive; TPSA: Topological polar surface area; t-SNE: T-distributed stochastic neighbor embedding.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13321-022-00619-2>.

**Additional file 1: Table S1** Model performance on threefold cross validations using M-data. **Table S2** Model performance on S-data trained by M-data with 5 times runs. **Table S3** Substrates in S-data. **Table S4** Model performance on substrates in S-data. **Table S5** Model performance of all combinations of uncertainty estimation methods in GROVER-BBBp model. **Table S6** A list of antitumor SMLs and corresponding prediction results in GROVER-BBBp model. **Table S7** Prediction results for external dataset collected from Colclough and Kim's works in GROVER-BBBp model. **Table S8** The details of 9 drug-like datasets from admetSAR. **Table S9** Model performance on fivefold cross-validations of 9 drug-like datasets in GROVER model. **Table S10** Model performance on test set of 9 drug-like datasets in GROVER model. **Fig. S1** Prediction performance by introducing different uncertainty estimation methods in 9 GROVER models. **Table S11** Model performance of various combination of uncertainty estimation methods on fathead minnow toxicity dataset in GROVER model. **Table S12** Model performance of various combination of uncertainty estimation methods on tetrahymena pyriformis toxicity dataset in GROVER model. **Table S13** Model performance of various combination of uncertainty estimation methods on AMES mutagenicity dataset in GROVER model. **Table S14** Model performance of various combination of uncertainty estimation methods on hERG inhibitor (II) dataset in GROVER model. **Table S15** Model performance of various combination of uncertainty estimation methods on CYP3A4 substrates dataset in GROVER model. **Table S16** Model performance of various combination of uncertainty estimation

methods on CYP1A2 inhibitor dataset in GROVER model. **Table S17** Model performance of various combination of uncertainty estimation methods on Caco-2 permeability dataset in GROVER model. **Table S18** Model performance of various combination of uncertainty estimation methods on P-gp inhibitor (I) dataset in GROVER model. **Table S19** Model performance of various combination of uncertainty estimation methods on P-gp inhibitor (II) dataset in GROVER model.

#### Acknowledgements

None.

#### Author contributions

XL and MZ directed the project. XT wrote the manuscript with the assistance of DW, XD, XT, QR, GC, YR, TX, JH and HJ. All authors read and approved the final manuscript.

#### Funding

This work was supported by Shanghai Municipal Science and Technology Major Project, the Lingang Laboratory (No. LG202102-01–02) and Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202002).

#### Availability of data and materials

All data and code to build the models are provided at: [https://github.com/tongxiaochu/BBB\\_uncertainty\\_project](https://github.com/tongxiaochu/BBB_uncertainty_project).

#### Declarations

#### Competing interests

The authors declare no competing interests.

#### Author details

<sup>1</sup>Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China. <sup>2</sup>University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China. <sup>3</sup>Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China. <sup>4</sup>School of Pharmaceutical Science and Technology, Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310024, China. <sup>5</sup>Tencent AI Lab, Shenzhen 518057, China.

Received: 17 September 2021 Accepted: 28 May 2022

Published online: 07 July 2022

#### References

- Di L, Rong H, Feng B (2013) Demystifying brain penetration in central nervous system drug discovery. *J Med Chem* 56:2–12
- Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 3:711–716
- Colclough N, Chen K, Johnstrom P, Strittmatter N, Yan Y, Wrigley GL, Schou M, Goodwin R, Varnas K, Adua SJ et al (2021) Preclinical comparison of the blood–brain barrier permeability of osimertinib with other EGFR TKIs. *Clin Cancer Res* 27:189–201
- Brown PD, Ahluwalia MS, Khan OH, Asher AL, Wefel JS, Gondi V (2017) Whole-brain radiotherapy for brain metastases: evolution or revolution? *J Clin Oncol* 36:483–491
- Patel NC (2020) Methods to optimize CNS exposure of drug candidates. *Bioorg Med Chem Lett* 30:127503
- Morales JF, Montoto SS, Fagioli P, Ruiz ME (2017) Current state and future perspectives in QSAR models to predict blood–brain barrier penetration in central nervous system drug R&D. *Mini-Rev Med Chem* 17:247–257
- Gabathuler R (2010) Approaches to transport therapeutic drugs across the blood–brain barrier to treat brain diseases. *Neurobiol Dis* 37:48–57
- Yu H, Yu Z, Jiang W, Hong L (2014) Lead compound optimization strategy (4)—improving blood–brain barrier permeability through structural modification. *Acta Pharm Sin* 49:789–799

9. Levin VA (1980) Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability. *J Med Chem* 23:682–684
10. Lobell M, Molnár L, Keserü GM (2003) Recent advances in the prediction of blood-brain partitioning from molecular structure. *J Pharm Sci* 92:360–370
11. Norinder U, Haeblerlein M (2002) Computational approaches to the prediction of the blood-brain distribution. *Adv Drug Delivery Rev* 54:291–313
12. Goodwin JT, Clark DE (2005) In silico predictions of blood-brain barrier penetration: considerations to “Keep in Mind.” *J Pharmacol Exp Ther* 315:477–483
13. Cruciani G, Pastor M, Guba W (2000) VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur J Pharm Sci* 11:529–539
14. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J Chem Inf Model* 45:1376–1384
15. Adenot M, Lahana R (2004) Blood-brain barrier permeation models: discriminating between potential CNS and Non-CNS drugs including P-glycoprotein substrates. *J Chem Inf Comput Sci* 44:239–248
16. Zhao YH, Abraham MH, Ibrahim A, Fish PV, Cole S, Lewis ML, de Groot MJ, Reynolds DP (2007) Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. *J Chem Inf Model* 47:170–175
17. Gupta M, Lee HJ, Barden CJ, Weaver DF (2019) The blood-brain barrier (BBB) score. *J Med Chem* 62:9824–9836
18. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A (2008) QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm Res* 25:1902–1914
19. Yuan Y, Zheng F, Zhan CG (2018) Improved prediction of blood-brain barrier permeability through machine learning with combined use of molecular property-based descriptors and fingerprints. *AAPS J* 20:54
20. Shaker B, Yu MS, Song JS, Ahn S, Ryu JY, Oh KS, Na D (2021) LightBBB: computational prediction model of blood-brain-barrier penetration based on lightGBM. *Bioinformatics* 37:1135–1139
21. Alsenan SA, Al-Turaiki IM, Hafez AM (2020) Feature extraction methods in quantitative structure–activity relationship modeling: a comparative study. *IEEE Access* 8:78737–78752
22. Roy D, Hinge VK, Kovalenko A (2019) To pass or not to pass: predicting the blood-brain barrier permeability with the 3D-RISM-KH molecular solvation theory. *ACS Omega* 4:16774–16780
23. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for molecular machine learning. *Chem Sci* 9:513–530
24. Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H et al (2020) Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 63:8749–8760
25. Wang S, Li Z, Zhang S, Jiang M, Wang X, Wei Z (2020) Molecular property prediction based on a multichannel substructure graph. *IEEE Access* 8:18601–18614
26. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, Leskovec J (2019) Strategies for pre-training graph neural networks. [arXiv:1905.12265](https://arxiv.org/abs/1905.12265). [arXiv.org e-Print archive. https://arxiv.org/abs/1905.12265](https://arxiv.org/abs/1905.12265).
27. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J (2020) Self-supervised graph transformer on large-scale molecular data. *NIPS* 33:12559–12571 [arXiv:2007.02835](https://arxiv.org/abs/2007.02835). [arXiv.org e-Print archive. https://arxiv.org/abs/2007.02835](https://arxiv.org/abs/2007.02835).
28. Christos EP, Hoi Y (2001) Uncertainty estimation and monte carlo simulation method. *Flow Meas Instrum* 12:291–298
29. Yu J, Li X, Zheng M (2021) Current status of active learning for drug discovery. *Artif Intell Life Sci*. <https://doi.org/10.1016/j.jaillsci.2021.100023>
30. Wang D, Yu J, Chen L, Li X, Jiang H, Chen K, Zheng M, Luo X (2021) A hybrid framework for improving uncertainty quantification in deep learning-based QSAR regression modeling. *J Cheminform* 13:69
31. Ding X, Cui R, Yu J, Liu T, Zhu T, Wang D, Chang J, Fan Z, Liu X, Chen K et al (2021) Active learning for drug design: a case study on the plasma exposure of orally administered drugs. *J Med Chem* 64:16838–16853
32. Schwill S (2018) Entropy Analysis of Financial Time Series. [arXiv:1807.09423](https://arxiv.org/abs/1807.09423). [arXiv.org e-Print archive. https://arxiv.org/abs/1807.09423](https://arxiv.org/abs/1807.09423).
33. Balaji Lakshminarayanan, Alexander Pritzel, Blundell C (2016) Simple and scalable predictive uncertainty estimation using deep ensembles. [arXiv:1612.01474](https://arxiv.org/abs/1612.01474). [arXiv.org e-Print archive. https://arxiv.org/abs/1612.01474](https://arxiv.org/abs/1612.01474).
34. Kononenko I (1989) Bayesian Neural Networks. *Biol Cybern* 61:361–370
35. Gal Y (2016) Uncertainty in Deep Learning. PhD thesis. University of Cambridge.
36. Gal Y, Ghahramani Z (2016) Dropout as A Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *Proceedings of the 33rd International Conference on Machine Learning*, Vol 48, PMLR, pp 1050–1059.
37. Kwon Y, Won J-H, Kim BJ, Paik MC (2020) Uncertainty quantification using bayesian neural networks in classification: application to biomedical image segmentation. *Comput Stat Data Anal* 142:106816
38. Janet JP, Duan C, Yang T, Nandy A, Kulik HJ (2019) A Quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem Sci* 10:7913–7922
39. Martins IF, Teixeira AL, Pinheiro L, Falcao AO (2012) A Bayesian approach to in silico blood-brain barrier penetration modeling. *J Chem Inf Model* 52:1686–1697
40. Doniger S, Hofmann T, Yeh J (2002) Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J Comput Biol* 9:849–864
41. Chen L, Zeng W-M, Cai Y-D, Feng K-Y, Chou K-C (2012) Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS ONE* 7:e32524
42. Abraham MH, Ibrahim A, Zhao Y, Acree WE Jr (2006) A Data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. *J Pharm Sci* 95:2091–2100
43. Wang W, Kim MT, Sedykh A, Zhu H (2015) Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm Res* 32:3055–3065
44. Friden M, Winiwarter S, Jerndal G, Bengtsson O, Wan H, Bredberg U, Hammarlund-Udenaes M, Antonsson M (2009) Structure–brain exposure relationships in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids. *J Med Chem* 52:6233–6243
45. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magarinos MP, Mosquera JF, Mutowo P, Nowotka M et al (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 47:D930–D940
46. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754
47. Kumar R, Sharma A, Tiwari RK (2013) Can we predict blood brain barrier permeability of ligands using computational approaches? *Interdiscip Sci* 5:95–101
48. Schyman P, Liu R, Desai V, Wallqvist A (2017) vNN web server for ADMET predictions. *Front Pharmacol* 8:889
49. Shaikh N, Sharma M, Garg P (2017) Selective fusion of heterogeneous classifiers for predicting substrates of membrane transporters. *J Chem Inf Model* 57:594–607
50. Wang X, Zhu X, Ye M, Wang Y, Li CD, Xiong Y, Wei DQ (2019) STS-NLSP: a network-based label space partition method for predicting the specificity of membrane transporter substrates using a hybrid feature of structural and semantic similarity. *Front Bioeng Biotechnol* 7:306
51. Cheng F, Li W, Zhou Y, Shen J, Wu Z, Liu G, Lee PW, Tang Y (2012) admet-SAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 52:3099–3105
52. Deane R, Singh I, Sagare AP, Bell RD, Ross NT, LaRue B, Love R, Perry S, Paquette N, Deane RJ et al (2012) A Multimodal RAGE-specific inhibitor reduces amyloid beta-mediated brain disorder in a mouse model of alzheimer disease. *J Clin Invest* 122:1377–1392
53. Berk C, Paul G, Sabbagh M (2014) Investigational drugs in alzheimer's disease: current progress. *Expert Opin Invest Drugs* 23:837–846
54. Green RC, Schneider LS, Amato DA, Beelen AP, Wilcock G, Swabb EA, Zavitz KH, Group TPS (2009) Effect of tarenfluril on cognitive decline and activities of daily living in patients with mild alzheimer disease: a randomized controlled trial. *JAMA* 302:2557–2564
55. Parepally JM, Mandula H, Smith QR (2006) Brain uptake of nonsteroidal anti-inflammatory drugs: ibuprofen, flurbiprofen, and indomethacin. *Pharm Res* 23:873–881
56. Eriksen JL, Sagi SA, Smith TE, Weggen S, Das P, McLendon DC, Ozols VV, Jessing KW, Zavitz KH, Koo EH et al (2003) NSAIDs and enantiomers of

- flurbiprofen target gamma-secretase and lower abeta 42 in vivo. *J Clin Invest* 112:440–449
57. Sun K, Mikule K, Wang Z, Poon G, Vaidyanathan A, Smith G, Zhang ZY, Hanke J, Ramaswamy S, Wang J (2018) A Comparative pharmacokinetic study of PARP inhibitors demonstrates favorable properties for niraparib efficacy in preclinical tumor models. *Oncotarget* 9:37080–37096
58. PharmaPendium <https://www.pharmapendium.com>. Accessed 28 May 2021
59. Mikule K, Wilcoxon K (2015) Abstract B168: the PARP inhibitor, niraparib, crosses the blood brain barrier in rodents and is efficacious in A BRCA2-mutant intracranial tumor model. *AACR 14:Abstract nr B168*
60. Shaw AT, Gandhi L, Gadgeel S, Riely GJ, Cetnar J, West H, Camidge DR, Socinski MA, Chiappori A, Mekhail T (2016) Alectinib in ALK-positive, crizotinib-resistant, non-small-cell lung cancer: a single-group, multicentre, phase 2 trial. *Lancet Oncol* 17:234–242
61. Ou S-H, Ahn JS, De Petris L, Govindan R, Yang JC-H, Hughes B, Lena H, Moro-Sibilot D, Bearz A, Ramirez SV (2016) Alectinib in crizotinib-refractory ALK-rearranged non-small-cell lung cancer: a phase II global study. *J Clin Oncol* 34:661–668
62. Lockney NA, Wu AJ (2017) Alectinib for the management of ALK-positive non-small cell lung cancer brain metastases. *J Thorac Dis* 9:E152–E154
63. Kodama T, Hasegawa M, Takanashi K, Sakurai Y, Kondoh O, Sakamoto H (2014) Antitumor activity of the selective ALK inhibitor alectinib in models of intracranial metastases. *Cancer Chemother Pharmacol* 74:1023–1028
64. Wang J, Gan C, Sparidans RW, Wagenaar E, van Hoppe S, Beijnen JH, Schinkel AH (2018) P-Glycoprotein (MDR1/ABCB1) and breast cancer resistance protein (BCRP/ABCG2) affect brain accumulation and intestinal disposition of encorafenib in mice. *Pharmacol Res* 129:414–423
65. Carr MJ, Sun J, Eroglu Z, Zager JS (2020) An evaluation of encorafenib for the treatment of melanoma. *Expert Opin Pharmacother* 21:155–161
66. Kim M, Laramy JK, Mohammad AS, Talele S, Fisher J, Sarkaria JN, Elmquist WF (2019) Brain distribution of a panel of epidermal growth factor receptor inhibitors using cassette dosing in wild-type and Abcb1/Abcg2-deficient mice. *Drug Metab Dispos* 47:393–404
67. Ameku K, Higa M (2020) Complete remission of multiple brain metastases in a patient with EGFR-mutated non-small-cell lung cancer treated with first-line osimertinib without radiotherapy. *Case Rep Oncol Med* 2020:9076168
68. Choo EF, Belvin M, Boggs J, Deng Y, Hoeflich KP, Ly J, Merchant M, Orr C, Plise E, Robarge K et al (2012) Preclinical disposition of GDC-0973 and prospective and retrospective analysis of human dose and efficacy predictions. *Drug Metab Dispos* 40:919–927
69. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A (2008) Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 48:1733–1746

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

