Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

Explaining Blood–Brain Barrier Permeability of Small Molecules by Integrated Analysis of Different Transport Mechanisms

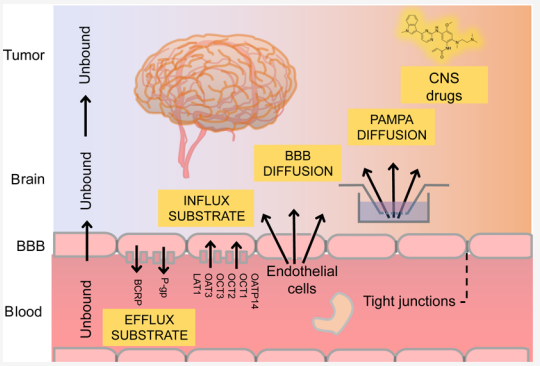
Fleur M.G. Cornelissen,[⊥] Greta Markert,[⊥] Ghislaine Deutsch, Maria Antonara, Noa Faaij, Imke Bartelink, David Noske, W. Peter Vandertop, Andreas Bender,* and Bart A. Westerman*

Cite This: *J. Med. Chem.* 2023, 66, 7253–7267

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The blood–brain barrier (BBB) represents a major obstacle to delivering drugs to the central nervous system (CNS), resulting in the lack of effective treatment for many CNS diseases including brain cancer. To accelerate CNS drug development, computational prediction models could save the time and effort needed for experimental evaluation. Here, we studied BBB permeability focusing on active transport (influx and efflux) as well as passive diffusion using previously published and self-curated data sets. We created prediction models based on physicochemical properties, molecular substructures, or their combination to understand which mechanisms contribute to BBB permeability. Our results show that features that predicted passive diffusion over membranes overlap with features that explain endothelial permeation of approved CNS-active drugs. We also identified physical properties and molecular substructures that positively or negatively predicted BBB transport. These findings provide guidance toward identifying BBB-permeable compounds by optimally matching physicochemical and molecular properties to BBB transport mechanisms.



INTRODUCTION

The central nervous system (CNS) is vulnerable to various diseases, such as brain tumors, Alzheimer's disease, autism, and multiple sclerosis. Therefore, CNS drugs comprise a wide range of chemotypes that cover various pharmacological categories. For these drugs to be effective, they must pass the blood–brain barrier (BBB) to reach a therapeutically relevant concentration in the brain. This is a major challenge in developing CNS drugs[2−4] as an estimated minority (only 2–6%) of all small compounds can enter the BBB, the "gatekeeper" of the CNS.[5−7]

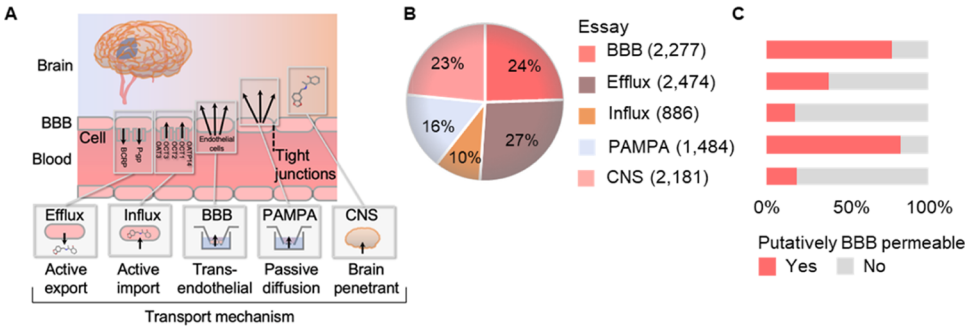
The brain is protected from much of the potential chemical interference by the BBB, an evolutionarily conserved barrier that selectively separates the brain from the circulatory system. The BBB consists of brain endothelial cells, their surrounding cells (smooth muscles, pericytes, astrocytes, and neurons), and a basement membrane.[8] The BBB controls the transport of ions, proteins, hormones, and immune cells between the brain and the blood, essential for adequate brain function.[8] These can be transported by either passive diffusion via intercellular tight junctions, closely connecting endothelial cells, or active transport by specific transporters on the surface of endothelial cells. Therefore, the BBB is one of the main hurdles of drug design for effective systemic treatment against CNS diseases.

Compounds can cross the BBB via passive transport either transcellularly through the phospholipid bilayer of the membranes of the BBB endothelial cells or paracellularly between endothelial cells where tight junctions restrict molecule diffusion. The efficiency of diffusion is dependent on physicochemical properties, such as lipid solubility and molecular weight.[5] This efficiency can be measured in vitro through, e.g., a parallel artificial membrane permeability assay (PAMPA).[9] In addition, endothelial cell culture models of brain uptake in vitro[10−12] are suitable for screening procedures with the BBB-specific expression of endothelial transporters and carrier proteins, to some degree imitating the in vivo situation.

For active transport, BBB passage is regulated by the transmembrane influx and efflux transporters.[13] Most influx transporters belong to the gene superfamily of solute carriers (SLCs).[14] SLC transporters can transport molecules of a wide range, including ions, neurotransmitters, and amino acids and, thus, are interesting targets for drug development.[15] There are a few drugs on the market binding specifically to SLCs in diseases.[16,17] The transporters belonging to *SLC22* and *SLCO* (formerly called *SLC21A*) are especially important for drug uptake including drug absorption, distribution, and elimina-

Received: November 17, 2022
Published: May 22, 2023

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253–7267



D

Dataset/Essay	Target(s)	Putative mechanism	Distribution outcome (measured transport effect)	Sources
BBB	BBB Membrane	BBB permeability	BBB permeable: 1,769 BBB impermeable: 508	Wang *et al.* 2018
Efflux	ABC Transporters	Efflux from brain	ABC substrate: 1,515 ABC non-substrate: 959	Metrabase (Mak *et al.*, 2015), ChEMBL (Mendez *et al.*, 2019)
Influx	SLC Transporters	Influx to brain	SLC substrate: 157 SLC non-substrate: 729	Metrabase (Mak *et al.*, 2015), ChEMBL (Mendez *et al.*, 2019)
PAMPA	Artificial Membranes	Passive membrane permeability	PAMPA permeable: 1,235 PAMPA impermeable: 249	ChEMBL (Mendez *et al.*, 2019)
CNS (Benchmark)	CNS Drug	BBB permeability	CNS drug: 411 Non-CNS drug: 1,770	Drugbank (Wishart *et al.*, 2018)

Figure 1. Overview of datasets used in their relation to putative BBB permeability. (A) Cartoon showing a schematic overview of different active and passive transport mechanisms for compounds to permeate the BBB. (B) Pie chart showing a comparison of dataset sizes based on the number of compounds within each dataset. (C) Histogram showing the corresponding overview of percentages of compounds assumed to be in favor of BBB permeation (red) or not in favor of BBB permeation (gray). (D) Table showing an overview of datasets with their sources used to generate the prediction models and their specific transport mechanism based on in vitro measurements (i.e., BBB, Efflux, Influx, and PAMPA) or based on observed clinical data (i.e., CNS).

tion.[18,19] These two subfamilies are particularly complex because they are not exclusively responsible for the influx of molecules but also the efflux of molecules as they facilitate transport in a bidirectional manner.[18,19] Still, SLCs are considered to be understudied.[16,17] Due to their complex structure, they are difficult to express and purify, which represents limitations to their experimental study. In addition, their biochemical, biophysical, and structural characterization is experimentally challenging.[16]

Similar to the SLCs, the superfamily of ATP-binding cassette (ABC) transporters has a wide range of substrate molecules. ABC transporters are—in contrast to SLCs—mostly entry-prevention transporters that recognize xenobiotics coming from the circulatory system and prevent them from passing the BBB to protect the brain. According to their evolutionary divergence, the ABC transporters are divided into seven subfamilies (ABCA−ABCG),[20] with *P*-glycoprotein (*P*-gp, ABCB1) as the most commonly reported active efflux system for small molecules.[21−23] *P*-gp plays a crucial role in drug resistance against antineoplastic and anticancer drugs, e.g., in colorectal cancer.[24]

Endothelial cell culture models of brain uptake in vitro are suitable for screening procedures with the BBB-specific expression of endothelial transporters and carrier proteins, to some degree mimicking the in vivo situation. A better representation of BBB transfer is the use of in vivo methods. This is commonly assessed in rodents by determining the unbound plasma/unbound brain concentration ratio of compounds[25] which is determined after giving the drug through its normal administration route.[26] Nevertheless, these data are only sparsely available in the public domain.

Over the past decades, absorption, distribution, metabolism, and excretion (ADME) property evaluation in laboratories has become one of the most important areas of interest in the process of drug discovery and development and has been widely accepted.[27] However, experimental evaluation and optimization with numerous in vitro assay methods are time-consuming and expensive,[28] and *in vitro-to-in-vivo* (IVIVE) usually contains a number of assumptions that increase the uncertainty of the estimated values. Therefore, alternative methods with lower costs such as computational models are needed for drug development and research. Nowadays, due to the high-speed development of computational technology, various in silico prediction models to evaluate compound ADME properties have been developed,[29−31] also for BBB transfer[32−36] (shown in Table S1).

In this work, we propose a new and more refined approach to assess BBB permeability by using previously published and self-curated data sets focusing on different aspects of BBB transport, respectively, via influx, efflux, and passive diffusion and through an endothelial in vitro model (Figure 1A). By using an eXtreme Gradient Boosting[37] (XGBoost) prediction model, we identified which physicochemical properties and/or molecular substructures can best explain the transfer of compounds to the brain. In addition, we improved the performance of these prediction models by combining physicochemical properties and molecular fingerprint data. After establishing the most optimal prediction models for each type of transport, we assessed their putative contribution to BBB permeation by providing them with an independent benchmark dataset consisting of approved small-molecule CNS drugs as well as non-CNS drugs. This showed that endothelial assays most accurately resembled CNS drug permeability and

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

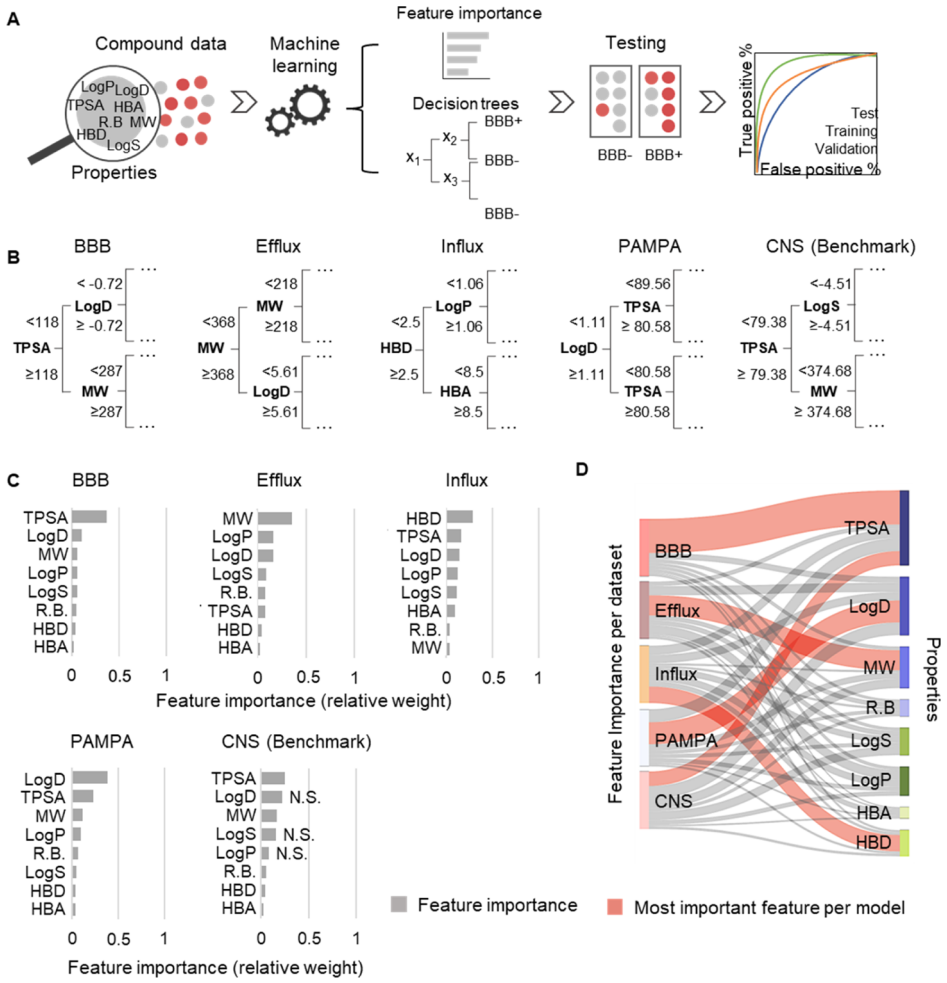


Figure 2. Workflow of the prediction models based on physicochemical properties per transport mechanism. (A) Schematic overview of the experiment. Physicochemical properties were used to train all five models separately, resulting in a ranked list of properties important for predicting the outcome that were tested on test and validation data within the same dataset. BBB+, BBB-permeable, BBB−, and BBB-impermeable. (B) Decision trees showing the first three decision splits per dataset. Cutoff values per property, defined by the outcomes from the XGBoost models, are depicted in each branch. (C) Histogram showing the relative feature importance of property-based models per dataset. Pearson's chi-squared test to rule out false positivity because of the overrepresentation of properties within each subset was performed (p value < 0.05; computed by Monte Carlo simulation). N.S., nonsignificant; MW, molecular weight; R.B., rotatable bonds; TPSA, topological polar surface area; Log*P*, partition coefficient of a molecule between aqueous and lipophilic phases; Log*S*, water solubility; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; Log*D*, log of partition of a chemical compound between the lipid and aqueous phases; (D) Sankey plot showing an overview of the relative feature importance (0−1) per dataset, with the most important physiochemical property feature per dataset depicted in red.

that the sum of BBB permeability can be caused by exclusive mechanisms, whereof passive transport appeared to contribute most to the net effect.

RESULTS

Compounds Can Be Transported via Various Transport Mechanisms into the BBB. The BBB is formed by specialized tight junctions between endothelial cells that line brain capillaries to create both a physical and selective biological barrier between the brain and the rest of the body, to exclude toxins from the brain. This exclusion also applies to most small-molecule compounds, limiting their bioavailability in the brain against CNS diseases such as brain cancer. Here, we investigated different transport mechanisms involved in the BBB permeation of small molecules to identify features that predict positive BBB transport. Different transport mechanisms are summarized in Figure 1A and details regarding the datasets used are provided in Figure 1B−D, with a total of 8658 unique compounds for which the transport mechanism has been determined, either quantitatively or qualitatively (Figure 1C).

Several compounds were shared between datasets (up to 7% between the CNS and Efflux), i.e., representing a minority of the cases (Figure S1A). The BBB dataset was used to investigate BBB permeation by an in vitro-grown CNS endothelial monolayer mimicking the BBB, wherein compounds can pass the endothelial cells via diffusion, influx, or efflux transporters. The PAMPA dataset is a strongly simplified representation of the BBB looking only at transcellular diffusion by using the PAMPA. The Influx dataset focuses on SLC transporters present in CNS endothelial cells, where specific compounds with SLC substrates can bind to pass the BBB. The Efflux dataset focuses on ABC transporters, also present in endothelial cells where compounds with ABC substrates can bind, resulting in the efflux of a compound out of the brain. Finally, the CNS dataset consists of approved small-molecular compounds classified as CNS and non-CNS drugs. According to our binary classification, the BBB and

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

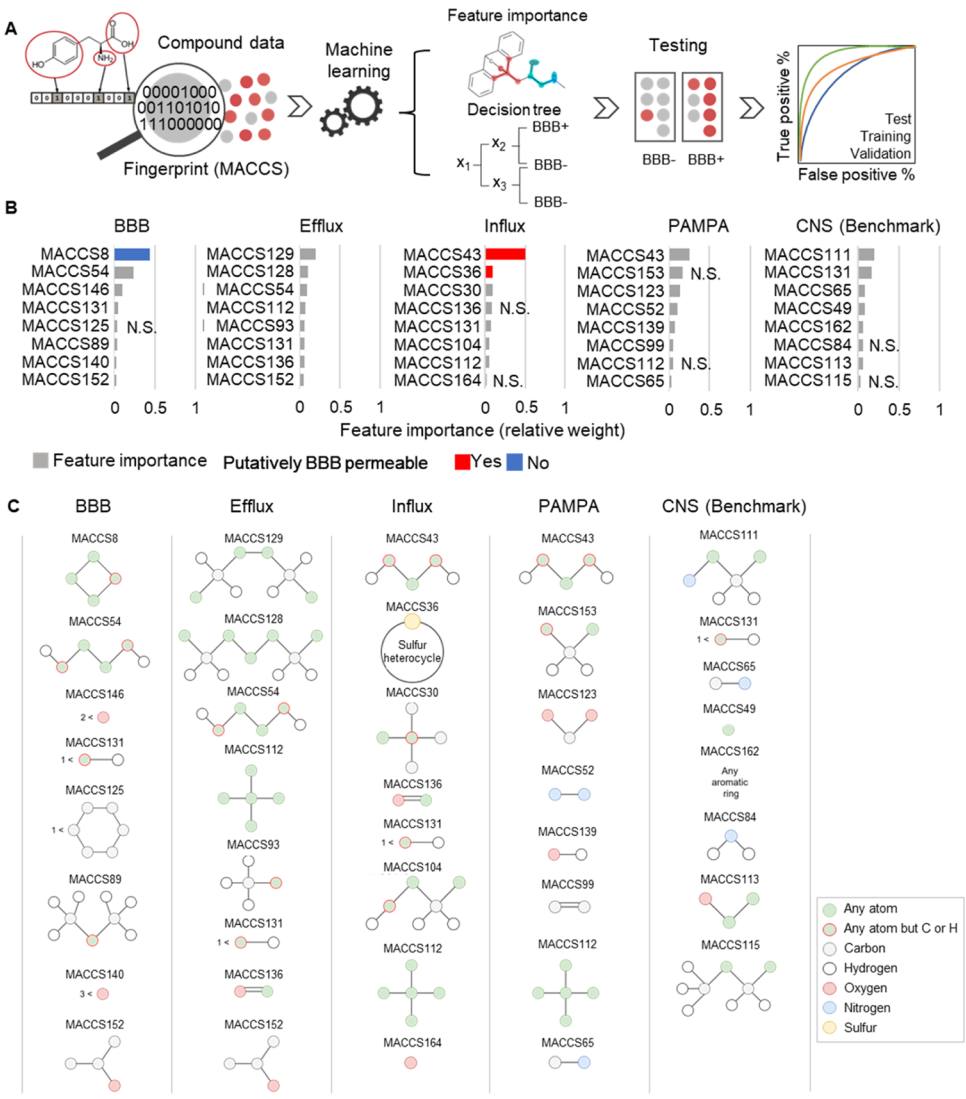


Figure 3. Workflow and results of prediction models based on the structural features per transport mechanism. (A) Schematic overview of the experiment. MACCS key fingerprints were used to train all five models separately, resulting in a list of molecular structures (MACCS keys) important for predicting the outcome that were tested on test and validation data within the same dataset. (B) Histogram showing the top 8 relative feature importance of MACCs key fingerprint models per dataset. Three examples of, respectively, BBB-permeable (red) and BBB-impermeable (blue) MACCS keys are highlighted and explained in (C). (C) Molecular substructures listed in Figure 3B. Pearson's chi-squared test was used to rule out false positivity because of the overrepresentation of fingerprints within each subset (p value <0.05; computed by Monte Carlo simulation). N.S., nonsignificant.

PAMPA datasets contain particular transport substrates in favor of putative BBB permeability (respectively, 78 and 83%; Figure 1B,C), whereas Efflux, Influx, and CNS datasets contain a minor group of substrates that would favor putative BBB permeability (respectively, 39 18 and 19%; Figure 1B,C).

t-distributed stochastic neighbor embedding (tSNE) analysis allows visualization of the chemical space based either on physicochemical properties [i.e., LogP and LogD, HBA, TPSA, rotatable bonds (RBs), molecular weight (MW), HBD, and LogS] or on molecular fingerprints (i.e., Molecular ACCess System (MACCS) molecular features; Figure S1B,C, respectively). Both properties and fingerprints show a spatial distinction in putative BBB permeability, whereas MACCS fingerprints show a slightly more diffuse profile. This indicates that the binary BBB permeation label does not directly correspond to fixed sets of properties/fingerprints but rather an assembly of multiple combinations that might be interdependent. This therefore urges us to analyze each transport mechanism separately. Indeed, when visualizing physicochemical properties and fingerprints per dataset (i.e., Efflux, Influx, PAMPA, and BBB) separately (Figure S1D,E), some distinct areas with varying BBB permeability can be observed.

BBB Transport Mechanisms Have Unique Physicochemical Profiles Essential for BBB Permeation. To investigate which physicochemical properties are important for BBB permeation per transport mechanism in our datasets, we developed an XGBoost prediction model for each dataset (Figure 2A, the performances of all models are shown in Table S1). XGBoost is a decision-tree-based ensemble machine learning algorithm, building a pre-set number of decision trees (e.g., classifiers) and creates a weighted combination of these trees, resulting in a prediction model where features can be ranked for importance on their outcome. Models were created on 65% of each dataset, and the remaining 35% was split into

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

an internal test and validation set to examine both the internal consistency and predictive ability of the model.

All prediction models resulted in an accuracy on the training data of 85% or higher on the training data and an area under the curve (AUC) ranging from 93 to 99%. The accuracy of the prediction models is illustrated in Figure S2A−E. The results of the validation (81.4%) and test (82%) sets are inferior to performance on the training set (93,4%), which is to be expected; however, the difference is relatively small, so the model does not seem to be overtrained. Imbalances between the positive or negative BBB permeability status within a dataset could lead to misclassification by the prediction model, as well as the presence of novel chemistry which the model is less familiar with. For example, the BBB and PAMPA validation and test sets (Figure S2A,D) show a lower negative predictive value (NPV) due to the overrepresentation of positive outcome data (Figure 1B,C), thereby resulting in more compounds that are falsely classified as BBB-impermeable. The other datasets on the other hand (Efflux, Influx, and CNS) show a lower positive predicted value (PPV, Figure S2B,C,E) compared to the training dataset, which can be attributed to the high number of BBB-impermeable compounds (Figure 1B,C), resulting in more false positives.

For each model, a total of three decision trees were created, with the first three splits of the first decision tree illustrated in Figure 2B. As shown, the first and therefore most important splits differ per dataset, implying that the transport mechanisms are driven by different physicochemical properties. An overview of important features (i.e., physicochemical profile) ranked in descending order of the decision trees per dataset is illustrated in Figure 2C. For the BBB dataset (i.e., endothelial assay), topological polar surface area (TPSA) appears to be the most important feature, suggesting that the charged atomic space defines permeation through brain endothelial cells. Indeed, published data on BBB permeability models also showed TPSA as the most important predictor for BBB permeation.[38−40] Similarly, we found that TPSA is also the most crucial descriptor for small-drug molecules to cross the BBB into the CNS (CNS dataset, Figure 2C). Thresholds for the TPSA in the BBB and the CNS dataset are <118 Å2 and TPSA <79.38 Å2, respectively (Figure 2B). This is in agreement with previously published data of CNS-active drugs, respectively, TPSA <90 Å2.[41,42] Systematic differences between the datasets could underlie the different thresholds, i.e., since the CNS dataset contains approved small-drug molecules only and the BBB dataset contains a variety of compounds, including usually large natural products, the latter consequently having a higher average TPSA.[43]

Concerning the efflux prediction model (i.e., referring to ABC transporters), MW is the most important predictor, also found in previously published data.[44−46] It should be noted that a simple cutoff for MW is not a reliable filter for distinguishing substrates from nonsubstrates since other physicochemical properties can affect efflux transport as well.[47] For influx via SLC transporters, HBD was the most important predictive feature. This is in agreement with previous literature, where HBDs were found to be important for drug−transporter interactions.[48] Interestingly, the feature ranking of influx versus efflux features is grossly inverted (Figure 2C), which might reflect differences in the mechanism of action between these types of transport.

For passive diffusion via PAMPA membranes Log*D*, a measure of lipophilicity, is the most important predictor for permeation which corresponds to the composition of the PAMPAs lipid bilayer composition (Figure 2D). TPSA, Log*D*, MW, and HBD are the overall most important factors to predict putative BBB permeation, whereas the contributions of RBs, Log*S*, Log*P*, and HBA are low in the current model.

In summary, different physicochemical properties are associated with the prediction of different transport mechanisms, matching earlier observations.[49,50] Since the predictive feature profile of the BBB versus the CNS dataset is almost identical, this could indicate that the net effect of passive and active transport is defined by the sum of these features. As shown in Figure 2C, only in the case of the CNS dataset, Log*D*, Log*S*, and Log*P* are of minor importance in predicting the outcome (Figure 2C). The results showing the most important features per model are summarized in Figure 2D.

Specific Molecular Substructures of Compounds Are Related to Effective BBB Transport. Besides physicochemical parameters, the precise molecular structure of a compound determines BBB permeation, both with respect to passive diffusion through lipophilic and hydrophilic phases as well as via recognition by influx or efflux transporters. To investigate if we can identify molecular substructures contributing to BBB permeation, we developed XGBoost models for all datasets using molecular fingerprints (Figure 3A). Again, XGBoost allows the identification of feature importance for the molecular fingerprints, for which we used MACCS molecular descriptors. This allows to back-translate the corresponding molecular substructure of a particular compound to the respective transport mechanism. Therefore, we aimed to identify specific molecular substructures that are associated with BBB permeation.

Compounds were described through MACCS fingerprints, consisting of a bitstring of 166 keys, with a subset of MACCS keys present in our datasets (Figure S4F). MACCS fingerprints were used to train the model and evaluated on a validation and test set. The chemical space of MACCS fingerprints color-coded by the binary outcome (i.e., putatively BBB-permeable or -impermeable) per dataset is illustrated in Figure S1E, where molecularly defined groups for the BBB and Influx datasets can be visually identified. In the case of Efflux, PAMPA, and CNS datasets, differences are less apparent.

XGBoost prediction model performance per dataset is shown in Figure S3A−E. All prediction models resulted in an accuracy on the training data of 72% or higher and an AUC ranging from 78 to 92%. For the validation sets, the accuracy is 71% or higher, and the AUC is 69% or higher. The worst-performing prediction model is the Efflux dataset, with an accuracy of 71%, and the best-performing prediction model is the Influx dataset with an accuracy of 89%.

For the MACCS models, we used up to five decision tree splits. Due to the sparsity and binary format of the input data, the addition of extra splits had no effect on the model performance. With this, we underlined the most important substructure(s) that predict putative brain penetrance. Each considered substructure has one corresponding MACCS key. Top-8 ranked MACCS keys per dataset are provided in Figure 3B and illustrated in Figure 3C, with the corresponding decision trees outlined in Figure S5A−E. As shown in Figure 3B, several MACCS keys overlap between datasets to predict the outcome. Both BBB and influx have a feature that is of high importance (feature importance >0.40) in their respective prediction models, MACCS8, which is a four-membered heteroatom ring, and MACCS43, which is one of any atoms

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

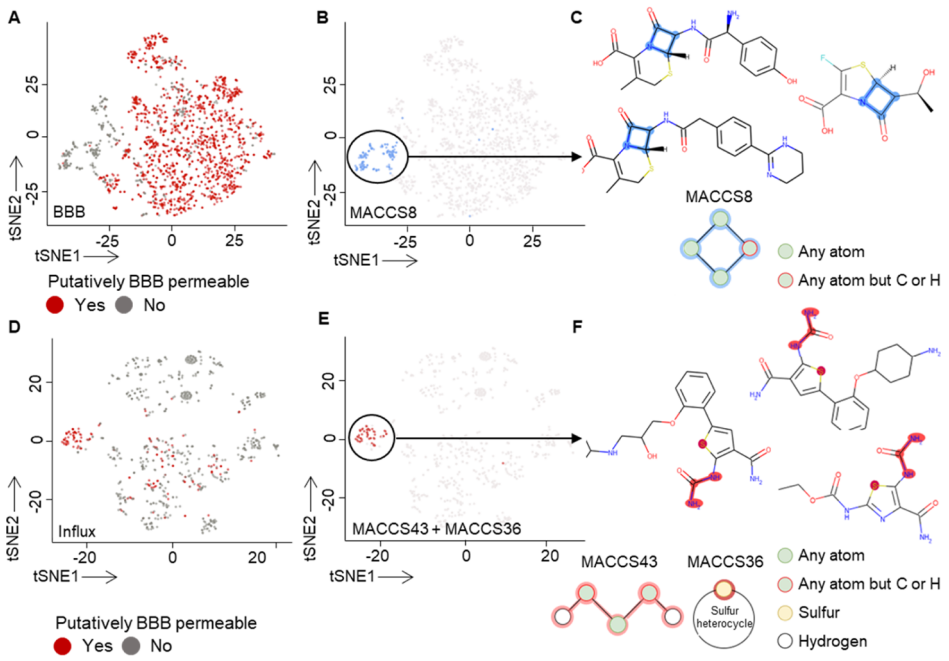


Figure 4. Specific molecular substructures predict putative BBB permeation and impermeation. (A) *t*SNE plot of compounds based on the MACCS key BBB dataset, with BBB-permeable compounds in red and BBB-impermeable compounds in gray. (B) *t*SNE plot of compounds of (C) with compounds containing MACCS 8, which is a four-membered ring with one heteroatom shown in blue. (C) Illustration of three compounds containing MACCS 8, a four-membered ring with one heteroatom of the BBB dataset, which are all BBB-impermeable. (D) *t*SNE plot of compounds based on the MACCS key Influx dataset, with putatively BBB-permeable compounds in red and BBB-impermeable compounds in gray. (E) *t*SNE plot of compounds of (F) with compounds containing MACCS43, which is one of any atoms bound to two other atoms except carbon or hydrogen, whereof each possesses one hydrogen, and MACCS36, a sulfur heterocycle shown in red, which are all putatively BBB-permeable. (F) Three examples of the structures of compounds containing MACCS43, which is one of any atoms bonded to two of the same atom each possessing one and only one hydrogen, and MACCS36, a sulfur heterocycle of the Influx dataset.

bound to two other atoms except carbon or hydrogen, whereof each possesses one hydrogen (Figure 3C). This is also supported by the fingerprint decision trees (Figure S5A–E), wherein both Influx and BBB datasets have only one short branch when following their respective MACCS8 and MACCS43 split (Figures S5A and S5C, respectively).

With the findings provided above, we focused on the interpretation of features from the BBB and Influx datasets since these datasets had similarities in feature importance (>0.4) of the most important features (Figure 3B). The tSNE based on the substructures of the BBB dataset shows a separation of BBB-impermeable compounds (gray; Figure 4A) and BBB-permeable compounds (red).

When highlighting all MACCS8-positive compounds, these all appear to be molecularly similar (Figure 4B) in the BBB-impermeable group (Figure 4A,B), suggesting that this particular four-membered ring with one heteroatom is of value in separating a large group of putative BBB-permeable versus -impermeable compounds in the dataset. This structure overlaps with the β lactam group, present in penicillin and its derivative antibiotics (158 out of 164 MACCS8-positive compounds, 96%), fitting the observation that β lactam-containing hydrophilic antibiotics are commonly excluded from the brain.[51,52] Consistently, the decision tree of the BBB MACCS model (Figure S5A) shows that the majority of MACCS8 positive compounds are BBB-impermeable, 154 (94%) impermeable and 10 (6%) BBB-permeable. Concerning the other branches of the decision tree, the separation of the model outcome is less apparent. To conclude, MACCS8 is associated with the exclusion of drugs from the brain, in the context of BBB permeability via endothelial cells. This is consistent with previous work, where a method using MACCS fingerprints and a support vector machine algorithm for BBB-permeable and -impermeable compounds found an important substructure for BBB-impermeable compounds.[53,54]

To independently validate the above-described findings, we performed the same tSNE analysis based on ECFP4 circular fingerprints, which captures higher levels of molecular complexity by describing more atom-to-atom relations within the molecule.[55] We used an ECFP bitstring of 1024 binary features. A tSNE visualization based on ECFP4 fingerprints for the BBB dataset is given in Figure 5A (left), wherein a similar pattern of putative BBB-permeable and impermeable compounds is found. Next, we took two subgroups of BBB-permeable and -impermeable compounds (Figure 5A, right) and determined the similarity of the compounds by calculating the Tanimoto similarity score, resulting in two major clusters (e.g., subgroups; Figure 5B).

Within these subgroups, compounds with a Tanimoto similarity score of ≥0.7 were analyzed in ChemMine Tools[56] to identify structural similarities between the compounds (Figure S6A). Similarities within subgroups A and B are shown in Figure 5C. A molecular substructure that was found earlier in the fingerprint analysis (i.e., MACCS8) was found again for the BBB-impermeable subgroup (Figure 5C, left figure; Figure S6A, subgroup A). BBB-permeable compounds were found to have a common structure consisting of aromatic rings (Figure 5C, right figure; Figure S6A, subgroup B1-3), consistent with earlier observations.[57,58] This latter group of molecules is similar to corticosteroids, known to enter the brain.[59] The BBB-impermeable compounds (Figure S6A, subgroup A)

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

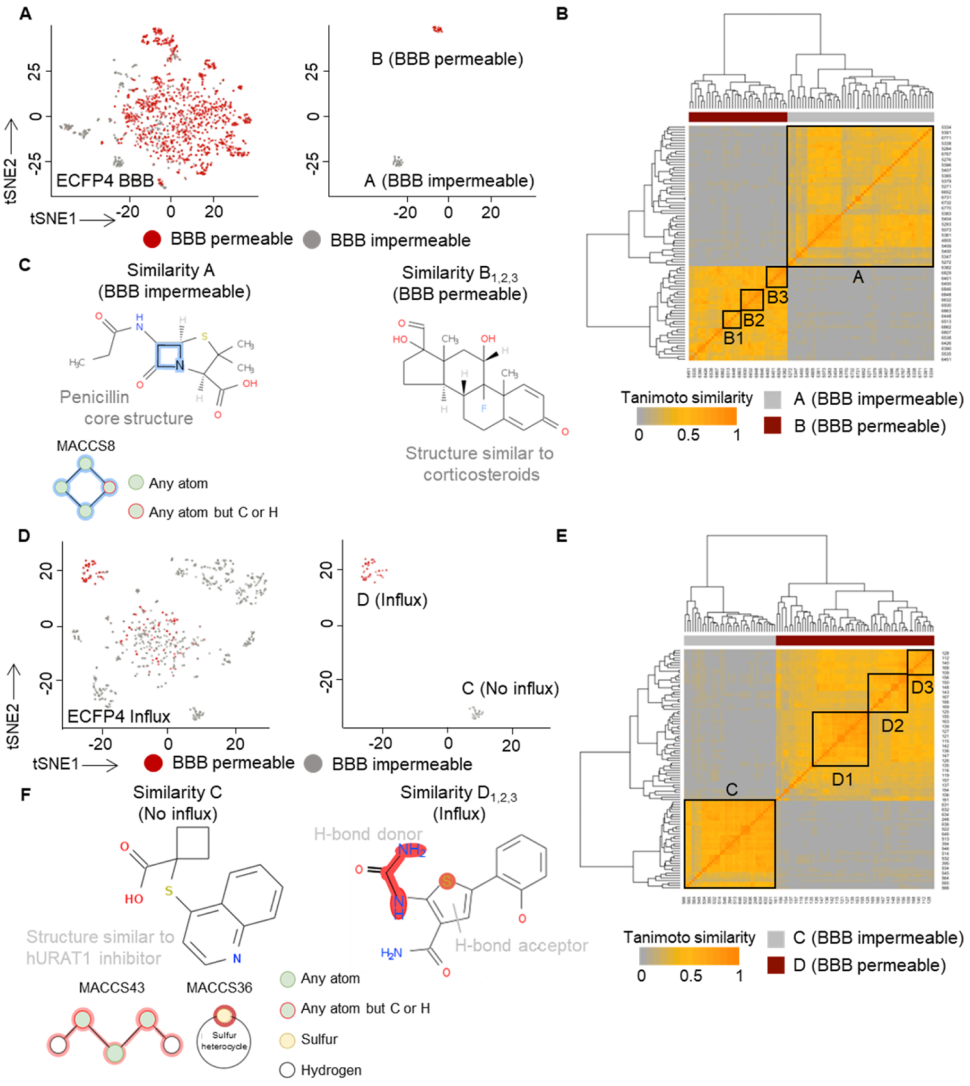


Figure 5. Prediction models based on ECFP4 fingerprints show identical results compared with MACCS fingerprint models. (A) tSNE plot of compounds of the BBB dataset based on ECFP4 fingerprints, with BBB-permeable compounds (red) and impermeable (gray). tSNE on the left shows a subset of the influx compounds, categorized into groups A (impermeable) and B (permeable). (B) Heat map showing Tanimoto similarity of compounds used in the BBB model, respectively, divided into groups A and B. For the molecular substructure search, group B was divided into three groups, respectively, B1, B2, and B3. For similarity search, one random compound of, respectively, group A, B1, B2, and B3 was taken and structures with a similarity of 0.7 or higher were used to find similarity structures between compounds. (C) Examples of compound-structure similarity for the BBB model for the impermeable group A (left) and permeable group B (right). The identified structure in group A contains the MACCS8 feature that was independently identified using the XGBoost BBB model based on MACCS keys (see Figure 3B,C). (D) tSNE plot of compounds present in the Influx dataset based on ECFP4 fingerprints, with BBB-permeable compounds (red) and impermeable (gray). tSNE on the left shows a subset of the influx compounds, categorized into groups A (impermeable) and B (permeable). (E) Heat map showing the Tanimoto similarity of, respectively, groups C and D. For molecular substructure search, group B was divided into three groups, respectively, B1, B2, and B3. For similarity search, one random compound of, respectively, groups C, D1, D2, and D3 was taken and structures with a similarity of 0.7 or higher were used to find similarity structures between compounds. Again, the substructure similarity for, respectively, compounds in group C (left) and group D (right). (F) As shown, a common structure that resembles hURAT1 inhibitors was found not to be associated with the influx. Interestingly, compounds positively associated with the influx contained common elements overlapping with substructures MACCS43 and MACCS36, previously identified using the XGBoost BBB model based on MACCS keys (see Figure 4D).

contain a beta-lactam structure, known to poorly permeate the BBB.[52,60]

Next influx (i.e., SLC substrate)-associated features were analyzed for Tanimoto similarity. The chemical space of influx data based on MACCS fingerprints is illustrated (Figure 5D), showing a group of compounds associated with no influx in gray (group C) and a group with influx shown in red (group D). Per subgroup, compounds with a Tanimoto similarity score of ≥0.7 (Figure 5E) were analyzed on the substructure level with ChemMine Tools. Similarities within the subgroups are shown in Figure 5F (all compounds are visualized in Figure S6B). Group C contains a structure similar to SLC22A12 solute carrier inhibitor61 hURAT1,[61,62] which might be explained through both being a transport inhibitor as well as a transport substrate. For influx permissive compounds, the same molecular substructures as identified with the MACCS XGBoost model were found, containing MACCS43 and MACCS36 (Figure 5F). MACCS43 contains two HBDs, correlating with the previous results seen in the property model. The chemical structure containing MACCS43 consists

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

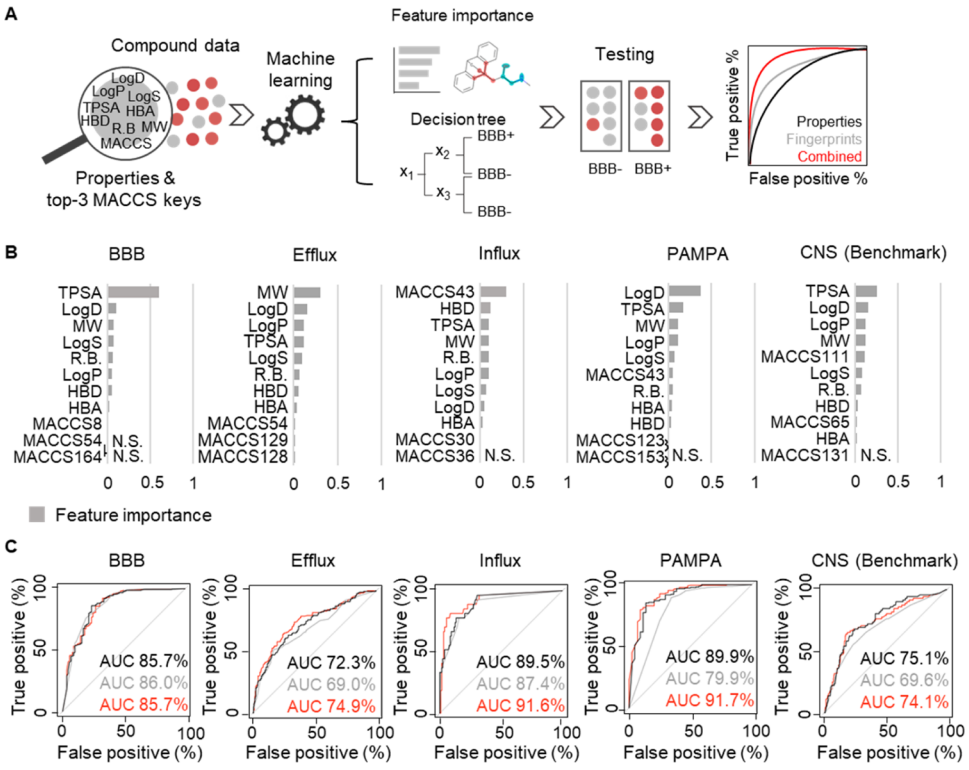


Figure 6. Combined prediction models of respective properties and the top three MACCS3 key fingerprints can improve prediction in comparison with properties only. (A) Schematic overview of the experiment. Physiochemical properties and the top three MACCS key fingerprints were used to train all five models separately, resulting in a list of features important for predicting the outcome that were tested on test and validation data within the same dataset. The results of the validation sets of both the XGBoost property model and fingerprint model and combined model were compared for prediction quality. (B) Histograms showing the relative feature importance of combined properties and the top three MACCS key fingerprint models per dataset. (C) Responder operator curve (ROC) containing, respectively, the properties, fingerprints, and combined AUCs of the validated datasets (i.e., BBB, Efflux, Influx, PAMPA, and CNS). For, respectively, the efflux, influx, and diffusion, the combination of properties and MACCS fingerprints in the model gives a slight improvement in the prediction compared with properties alone, whereas for the BBB and CNS, the prediction is similar/worse compared with properties alone. Pearson's chi-squared test was used to rule out false positivity because of the overrepresentation of properties or fingerprints within each subset (p value <0.05; computed by Monte Carlo simulation). N.S., nonsignificant.

in 90.6% (58/64 compounds) of the cases of two amino groups connected by a carbon and is directly attached to the MACCS36 structure, a sulfur heterocycle (in 98%, 63/64 of the compounds). In addition, the carbon atom of MACCS43 predominantly had a double bond with oxygen in 98% (63/64 compounds) of the cases. Of the MACCS43- and MACCS36-containing compounds, 64 are BBB-permeable (85%) and 11 are BBB-impermeable (15%; Figure S5C, including a detailed overview of these compounds on a molecular level in Figure S8). In conclusion, with this dataset, we found a large subset of SLC substrates that can be distinguished from SLC non-substrates by containing the MACCS43 and MACCS36 substructures in their molecule.

BBB Permeability Predictions Based on Properties and Molecular Substructures Can Improve Model Performance. Eventually, we created a final model per dataset with the physicochemical properties and the three highest-ranked MACCS keys combined. In addition, we compared the result for all three models individually for either properties or fingerprints and their combination (Figure 6A).

The combined models showed an accuracy ranging from 87 to 97% (Figure S4A–E). The test sets (i.e., accuracy 69–90%) and validation sets (i.e., accuracy 71–91%) gave less well-predicted outcomes compared with the training sets. For most datasets, e.g., BBB, Efflux, PAMPA, and CNS, the physicochemical properties are the most important features for predicting the permeability (Figure 6B). On the contrary, for the Influx dataset, the MACCS43 key was found to be of greater importance than the properties. When looking at the AUCs of the validation sets for the influx datasets of, respectively, the property model (87%), fingerprint model (87%), and combined model (91%), the combined model of influx resulted in a better prediction compared with properties only and the fingerprint model only (Figure 6C). This is in agreement with the feature importance list (Figure 6B), where the MACCS43 substructure is of greater importance than any of the properties, suggesting that molecular substructures are important for active or passive influx transport. Improved performance of the prediction models with properties and MACCS fingerprints combined was seen for Influx (from 87 to 91%), Efflux (from 70 to 72%), and PAMPA (from 88 to 90%) datasets, although the improvement was limited (Figure 6C). For both BBB (from 87 to 87%) and CNS (from 78 to 78%) datasets, adding fingerprints did not improve the models, probably since TPSA is a dominant predictor in both models overruling the relative weight of other predictive features.

Endothelial BBB Model Most Accurately Predicts the CNS Benchmark Data. In our final analysis, we aimed to determine the potential contribution of each model to the process of actual drug delivery to the brain. For this, we used

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

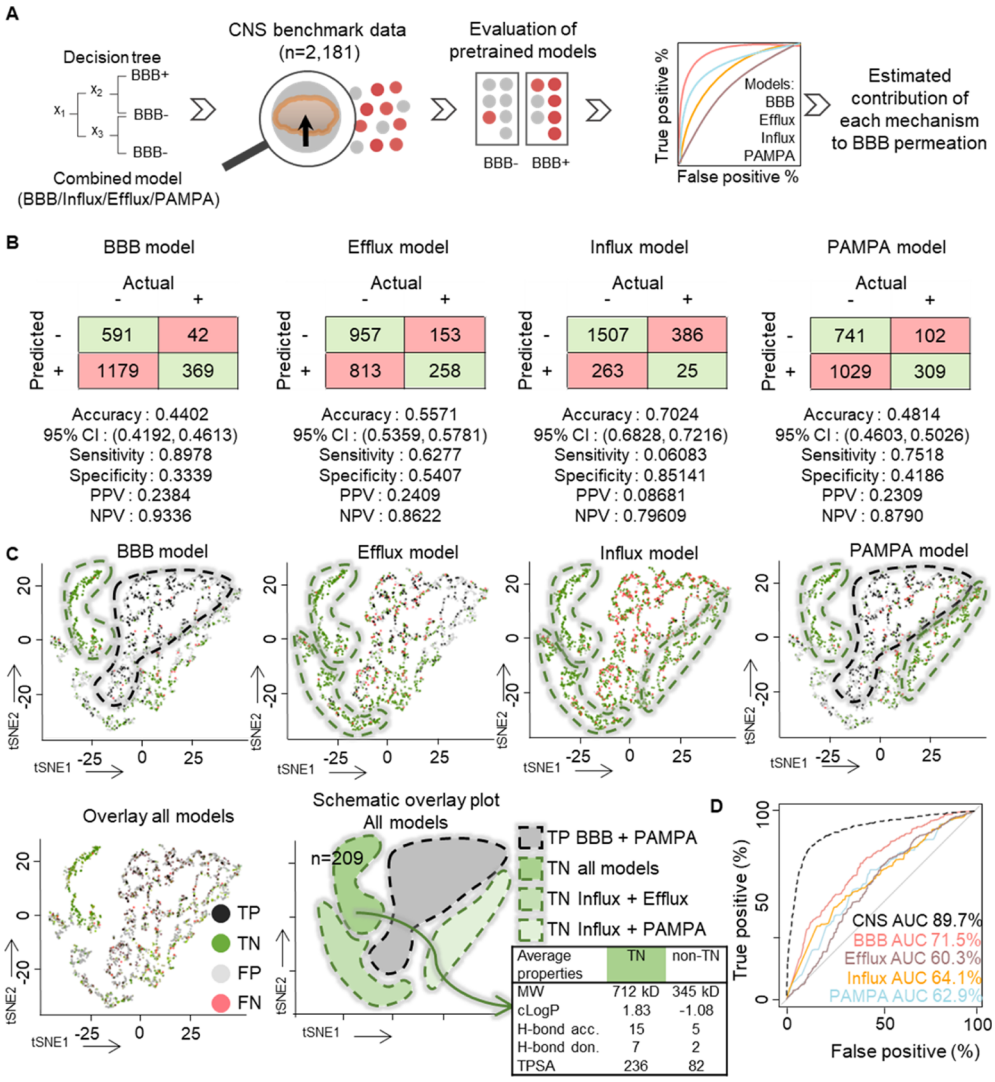


Figure 7. Benchmarking the individual BBB models using an unseen CNS benchmark dataset. (A) Schematic overview experiment. The combined models of, respectively, BBB, Influx, Efflux, and PAMPA were used to predict the CNS dataset. (B) Contingency table showing the overview results of predicted CNS data with, respectively, the BBB, Efflux, Influx, and PAMPA models. The BBB model appeared to predict the CNS drugs (i.e., drugs penetrating the brain) of the CNS dataset best. CI, confidence interval; PPV, positive predicted value; NPV, negative predicted value. (C) tSNE plot showing the overview of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predicted CNS compounds per prediction model (i.e., BBB, Efflux, Influx, and PAMPA) and an overlay of all prediction models. The schematic overlay plot shows the different areas that are either true-positively predicted (gray) or true-negatively predicted (shades of green). The average properties of TN (n = 209) and the non-TN drugs (n = 1987) are shown in the inset, showing that TN drugs have a higher MW, higher hydrophilicity, higher amounts of H-bond acceptors and donors, and a higher polar surface area (TPSA). See Figure S7 for details. Predicted. (D) ROC of CNS data predicted with, respectively, BBB, Influx, Efflux, and PAMPA models. The dashed black line represents the ROC curve of the CNS combined model on the CNS dataset (benchmark).

an experimental CNS benchmark dataset, whose compounds are known to have activity in the brain and hence need to reach this target organ via one of the mechanisms available for crossing the BBB. Again, we used the combined models (i.e., based on properties and top three MACCS keys) of, respectively, BBB, Influx, Efflux, and PAMPA (Figure 7A) and tested them on the unseen CNS data set.

Comparisons of the models are illustrated in Figure 7B. The level of correspondence of each drug of the CNS dataset is shown as a *t*SNE visualization in Figure 7C, also showing commonly observed patterns as seen over the different prediction models. Interestingly, a number of non-CNS-penetrating drugs were consistently predicted to be non-penetrant for all models (shown as dark green in the schematic overlay plot). These non-CNS-penetrating compounds (209 out of 1442 nonpenetrating drugs; 14%) have a significantly larger MW, hydrophilicity, number of HBDs and HBAs, as well as a larger TPSA, all matching previous reports[3] (average values shown in Figure 4C, and statistical analysis is shown in Figure S7). For the remaining compounds, different modes of action might apply to subgroups of compounds. This implies that BBB penetrance can for a large part be ascribed to a blend of different transport mechanisms. The BBB model resulted in the highest sensitivity (89.78%) with 258 out of 411 TPs (Figure 7B,C). Given the similarity between the results (Figure 7B,C) of the BBB and PAMPA datasets, this suggests that passive diffusion can account for an important part of the BBB permeability of CNS compounds, see also the schematic

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

overlay plot. Still, around 25% (102 out of 411) of the PAMPA-predicted CNS drugs are not identified as BBB-positive compounds, showing that other transport mechanisms play a role as well. In addition, most CNS compounds do not appear to be Efflux substrates as 258 compounds (63%) are correctly predicted as nonsubstrates with the Efflux model. The overlap between the models is also shown in Figure 7C, again highlighting that the most consistent true positive outcomes of the sum of the models resemble the BBB and PAMPA models.

The comparisons between the model performance can also be assessed through AUC analysis, showing that the BBB model has the best retrieval of CNS compounds at 71.5% compared with the other models (Figure 7D) and is closest to the AUC of the model of the CNS itself (89.7%, Figure 7D). This can in other words also be described as similarities between both machine learning models (and the underlying datasets) of both the CNS and BBB, containing a generally similar ranking of features important for prediction of their outcome (Figure 6B). Together, by evaluating the role of each transport model onto a benchmark CNS dataset, our data suggest that the BBB model most accurately predicts the CNS benchmark data and that all models together correctly predict compounds as non-CNS penetrant based on a large MW, hydrophilicity, proton donors and acceptors, as well as a high TPSA. When looking at each transport mechanism individually, i.e., passive diffusion, influx, and efflux, most CNS compounds seem to permeate through the BBB by passive diffusion, based on the data and analysis performed in this work.

DISCUSSION AND CONCLUSIONS

The pharmacological activity of CNS therapeutics depends on the ability to cross the BBB as well as on its availability for target engagement in the brain. Limited BBB permeation is one of the most important factors that limit the effectivity of CNS drugs,[5] and it restricts most repurposing options of non-CNS drugs for CNS indications such as brain tumors. Experimental models to assess BBB permeation for therapeutics are expensive and time-consuming, limiting the successful development of drugs against neurological diseases.[63] As an attractive alternative, computational prediction models could complement these efforts with a significantly lower cost. In this work, we were able for the first time to assemble datasets for different transport mechanisms, match them to BBB permeation measurements, and distinguish which parts of chemical space are likely to be translated to transport mechanisms to the brain. For this, we assembled several large, publicly available bioassay datasets from previously published as well as from self-curated data sets to enable a computational assessment of BBB barrier crossing based on different transport mechanisms. Our approach indicates that integrating all types of BBB transport in a one-fits-all prediction model is not likely to work since different features of compounds contribute to different transport mechanisms, with some exceptions. Instead, we think that an exclusive assessment of CNS transport mechanisms has more meaning and is consistent with the debate in the field, which shows complementary and sometimes conflicting contributions of different transport mechanisms.[13,49,64−67]

We identified several physicochemical and structural features, of which some have been identified before in the literature as being important for BBB permeation.[43−46,53,57,65] Interestingly, we found a submolecular group (i.e., MACCS43) that was top-ranking for both Influx and PAMPA models, favoring putative BBB permeation. The combination of this subgroup with MACCS36, also a top-ranking feature for the Influx model, might provide compounds with the possibility to act amphiphilically through conformationally masking HBDs consisting of two carbon-linked amino groups by a thiolane group, which has been identified as HBA (i.e., thiolane group in MACCS36),[68] similar to those observed in other contexts.[69] This new mechanism might open up several possibilities for the design of new compounds where HBDs and HBA are placed in close vicinity. Conformational states as proposed to exist here can translate to unexpected physical properties[70] which might help to identify such compounds.

In general, the relative contribution of features within each model could be skewed due to imbalances and complexity differences within and between datasets. To overcome this kind of bias, we tried to create sufficient statistical power by extracting compounds from large curated databases (e.g., ChEMBL, Metrabase, and Drugbank), although the increased power of these datasets comes with a cost due to possible biases that inherently arise upon their assembly. In addition, some compounds are only tested for one transporter, e.g., being an *SLC22* substrate for the Influx dataset, and therefore, not all *SLCO* substrates are represented. The substructure analysis depends on the existence of the respective fragments in the training set of a model; for instance, penicillin derivatives represented by MACCS8 are twofold more often present in the BBB dataset when compared to the other sets. To take potential disbalances into account, we assessed potential false positivity using a Chi2 test to identify major disbalances within datasets, which applied to some features shown as not significant (N.S.) in the histograms. Imbalances between the BBB and CNS datasets might account for a limited predictive value of 23.84% for the BBB prediction model based on physicochemical and structural features, which can be ascribed to the high number of BBB-permeable compounds in the BBB dataset compared with the low number of BBB-permeable compounds in the CNS dataset, resulting in a "skewed" model.

In conclusion, our study shows that different and sometimes exclusive transport mechanisms are expected to be involved in the BBB permeation of small-molecule drugs and which parts of physicochemical property and chemical substructure space are more likely to be involved in which type of mechanism depends on the respective context. All models correctly predicted non-CNS penetrance based on physicochemical properties such as size and charge. The individual model trained on data of transport across endothelial cells provided the most accurate BBB transfer model. This model showed the most resemblance to passive diffusion as compared to influx or even to a lesser extent, drug efflux mechanisms. Our assembled data, provided scripts, and models could therefore be used as relevant guidance to medicinal chemists, to discover potential BBB-permeating compounds across mechanisms and more easily rapidly reject or modify compounds with poor BBB permeability to become more effective in treating CNS diseases.

EXPERIMENTAL SECTION

Generation of Datasets. *BBB Dataset.* For the BBB dataset (i.e., endothelial essay), we used the biggest publicly available data set for BBB permeation.[34] The dataset combines four other data sets,[36,53,71,72] which are categorized as *BBB-permeable* (LogBB ≥

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

−1, with LogBB defined as the logarithmic ratio between the concentration of a compound in the brain and blood) and *BBB-nonpermeable* (LogBB < −1). Duplicates and self-contradictory compounds were removed, resulting in a dataset of 2277 compounds, containing, respectively, 1769 permeable and 508 impermeable compounds. The BBB dataset was used to investigate BBB permeation by an in vitro-grown CNS endothelial monolayer mimicking the BBB, wherein compounds can pass the endothelial cells via diffusion, influx, or efflux transporters.

Efflux Dataset. The Efflux dataset focuses on ABC transporters, also present in endothelial cells where compounds with ABC substrates can bind, resulting in the efflux of a compound out of the brain. For the Efflux dataset, we used previously published data[73] with human *P*-gp as a target, extracted from the ChEMBL database (version 25).[74] Based on the efflux ratio (ER), the compounds were classified into substrates (ER ≥ 5) and nonsubstrates (ER ≤ 1), and compounds with an ER between 1 and 5 were discarded. This stringent criterion was also used by Esposito et al. to obtain the most sufficient results.[73] Compounds that were tested in Caco-2 or MDCK cell lines were included, resulting in a binary data set of 1082 compounds with 701 substrates and 381 nonsubstrates.[75]

Additionally, data was extracted from Metrabase[75] containing information on *P*-gp as well as on ABC transporters located at the BBB. We chose the targets MDR1 (also known as *P*-gp; *ABCB1*) with 448 substrates and 355 nonsubstrates, BCRP1 (*ABCG2*) with 277 substrates and 131 nonsubstrates, MRP1 (*ABCC1*) with 57 substrates and 52 nonsubstrates, MRP2 (*ABCC2*) with 65 substrates and 53 nonsubstrates, MRP3 (*ABCC3*) with 23 substrates and 86 nonsubstrates, and MRP4 (*ABCC4*) with 18 substrates and 5 nonsubstrates, which are the efflux transporters located at the BBB.[76−78] This resulted in a total set of 1570 compounds, respectively, with 888 substrates and 682 nonsubstrates. The combined datasets with deletion of overlapping compounds in both datasets resulted in an Efflux dataset of 2474 compounds in total, whereof 1515 were substrates and 959 were nonsubstrates.

Influx Dataset. The Influx dataset focuses on SLC transporters that are present in CNS endothelial cells, where specific compounds with SLC substrates can bind to pass the BBB. Transporter data was collected from ChEMBL[74] and Metrabase.[75] In Metrabase, the keywords "SLC22" and "SLCO" resulted in 240 compounds, whereof 140 were classified as substrates and 100 as nonsubstrate. In the ChEMBL database, we searched for compounds from the subfamilies "SLC22" and "SLCO" and filtered for "*Homo sapiens*" as the target organism and "single protein" as the target type. Next, we performed text mining in the provided description of each bioassay using the keywords: "inhibition", "substrate", "binding affinity", "displacement", "TP transporter", "activity", "potency", "transport", and "induction", grouping together the biological activities of compounds which in their assay description include the same keyword. We manually examined the corresponding scientific publications describing the bioassays and defined if the biological activities are related to the permeation of the compounds in the transporter or not in order to classify the bioactivities into "uptake" and "nonuptake". The bioactivities were annotated as "active" based on the following criteria: pChEMBL >5 or bioactivity <10,000 nM or alternatively the comment section in ChEMBL containing the keywords "active", "permeable", "substrate", or "inhibitor". Data points with a pChEMBL <5, or a bioactivity >10,000 nM, or alternatively their comment section in ChEMBL containing the keywords "inactive" or "not active" were annotated as "inactive". In conclusion, the compounds with the corresponding SLC transporters were classified into four groups: "active uptake", "inactive uptake", "active nonuptake", and "inactive nonuptake". In order to achieve a binary classification of "substrate" and "nonsubstrate", "active uptake" was relabeled to "substrate" for 17 compounds because there is evidence that these compounds are substrates under a particular set of conditions, and the remaining categories were relabeled to "nonsubstrate" for 629 compounds. With now matching classifications, the data from Metrabase and ChEMBL were combined. Compounds with contradictory data from either the same subfamily were deleted, resulting in a data set with 886 compounds with a binary classification according to whether they are a substrate (157 compounds) or a nonsubstrate (729 compounds) of SLC22 or SLCO transporters.

PAMPA Dataset. The PAMPA dataset is a strongly simplified representation of the BBB looking only at passive transcellular diffusion by using the PAMPA.[79] Transcellular diffusion (i.e., passive membrane permeability) can be measured by usage of the PAMPA.[9] The test compound is placed in a container (i.e., donor well) that is connected to another solution without any test compound molecules present (i.e., acceptor well) via the artificial membrane where, after a specific time, the UV absorption of both the donor solution and the acceptor solution is measured to quantify the molecules that diffused from the donor well to the acceptor well.[34,80,81] The main drawback of cell-free methods is that PAMPA experiments can observe passive permeability only, neglecting special characteristics such as the active transporters acting at the BBB. For retrieving a data set containing PAMPA measurements, the ChEBML database was screened for entries containing the keyword "PAMPA". In the next step, these results were again filtered for the keywords "BBB", "brain", or "pH 7". Only those compounds were selected where information on the permeation rate Pe was included. Molecules with a Pe > 4×10^6 cm·s^{-1} were considered permeable, and whole molecules with a Pe < 2×10^6 cm × s^{-1} were considered nonpermeable.[34] All other compounds were discarded. This resulted in a dataset of 1484 compounds, with 1235 PAMPA permeable and 249 PAMPA impermeable compounds.

CNS Dataset. For the CNS (i.e., benchmark) dataset, we used Drugbank, an online database for drug molecules and approved drugs, both small molecules and biopharmaceuticals.[82] We extracted all approved small-molecular drugs and classified them according to the presence or absence of the label "central nervous system agent", resulting in a total of 2195 compounds, with 412 CNS compounds and 1783 non-CNS compounds.[1,84−87]

For subsequent analysis, all compounds were classified in binary outcomes, all relating to a putative BBB permeable status (e.g., BBB-permeable, influx substrate, efflux nonsubstrate, and CNS-active small molecules), a putative BBB impermeable status (e.g., BBB-impermeable, influx nonsubstrate, efflux substrate, and non-CNS-active small molecules).

Computation of Physicochemical Properties. Physicochemical properties such as the number of hydrogen bond donors (HBDs), number of hydrogen bond acceptors (HBAs), number of RBs, and MW were calculated using the RDKit package[83] in Python. ACD Labs Percepta (Advanced Chemistry Developments Inc., 2021, Toronto, ON, Canada, www.acdlabs.com) was used to predict *P*-gp probability, LogP, TPSA, polarizability, LogS (at pH 7.4), and LogD (at pH 7.4).

Fingerprints and Tanimoto Similarity Assessment. For the generation of MACCS fingerprints and extended-connectivity fingerprints 4 (ECFP4), the rcdk package84 version 3.6.0 in R (version 4.0.3) 85 was used. MACCS keys could be translated with the RDKit package86 in Python by the usage of a predefined dictionary (i.e., MDL keys87) containing a SMARTS list of substructure patterns. There was a one-to-one correspondence between each SMARTS pattern and the bit in the fingerprint. If the specified substructure was present in a compound, the corresponding bit was set to "1"; conversely, it was set to "0" once the substructure was absent in the compound. For the visualization of MACCS keys, by using the SMARTS patterns, the RDKit package in Python was used. To calculate Tanimoto's similarities of fingerprints, the fingerprint package version 3.5.7 in R was used. ChemMine Tools56 was used for the visualization of compound substructure similarities based on ECFP4.

XGBoost Models. For the generation of prediction models, the XGBoost package[37] version 1.5.0.2 in R was used. When top-predicting features were extracted, potential positively or negatively enriched features in the datasets were filtered using a chi-square test.

Computational Details. Datasets were split into a training (65% data), test (17.5% data), and validation (17.5% data) set. The sets were split by a data splitting function (caret package version 6.0)[88] to generate sets with equal binary outcomes (status 0 or 1). The training,

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

test, and validation sets for each dataset were identical for all three XGBoost models, namely, the property, fingerprint (MACCS), and combined (property and fingerprint) models.

For the XGBoost models based on properties, TPSA, MW, Log*D*, Log*S*, Log*P*, HBD, and HBA were used. The settings for the XGBoost function were a maximum depth of 7, a number of threads of 1, and a number of rounds of 3, which were found to be appropriate settings in preliminary model performance explorations. The model output ranged from 0 to 1, and a cutoff of ≥0.5 was used to classify the predictions into 0 (no BBB permeation) and 1 (BBB permeation).

For the XGBoost models based on fingerprints (MACCS keys or ECFP4), all 166 keys or 1024 bits were used as a predictor for the outcome. To identify the most important MACCS keys to predict BBB permeation per dataset, the XGBoost settings were set to a maximum depth of 5 and a number of rounds of 1.

For the combined XGBoost models, the property data in combination with the three most important MACCS keys based on feature importance (XGBoost package) were used. XGBoost model settings were identical to the settings used for the models based on properties to obtain the best comparability (i.e., a maximum depth of 7, a number of threads of 1, and a number of rounds of 3).

■ ASSOCIATED CONTENT

Data Availability Statement

All data and written scripts can be found at https://github.com/bartwesterman/Cornelissen-et-al

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c01824.

> Overview of similarities between datasets and chemical space analysis based on physiochemical properties and fingerprints per dataset, confusion matrices and corresponding AUCs of physiochemical property models, confusion matrices and corresponding AUCs of MACCS fingerprint models, confusion matrices and corresponding AUCs of physiochemical property and structural feature models, decision tree MACCS fingerprint models, compounds' Tanimoto similarity search, analysis of the CNS dataset for the TNs for all models, detailed MACCS structural analysis of the molecular influx model, and overview of the performance of all XGBoost models (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Andreas Bender − *Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*; orcid.org/0000-0002-6683-7546; Email: ab454@cam.ac.uk

Bart A. Westerman − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands; Window Consortium*, www.window-consortium.org; orcid.org/0000-0002-9898-9616; Email: a.westerman@amsterdamumc.nl

Authors

Fleur M.G. Cornelissen − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands*

Greta Markert − *Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*; orcid.org/0000-0001-5254-5596

Ghislaine Deutsch − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands; Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*

Maria Antonara − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands; Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*

Noa Faaij − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands*

Imke Bartelink − *Department of Pharmacy, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands*

David Noske − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands*

W. Peter Vandertop − *Department of Neurosurgery, Amsterdam UMC, Amsterdam 1105 AZ, the Netherlands*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.2c01824

Author Contributions

⊥F.M.G.C. and G.M. contributed equally.

Author Contributions

B.A.W., F.M.G.C., G.M., and A.B. developed the conceptual framework. G.M. and M.A. generated the datasets. F.M.G.C. and G.M. performed the bioinformatic analyses. G.D. and N.F. created fingerprint structure figures. B.A.W., A.B., W.P.V., and D.N. enabled the project by providing resources and supervision. F.M.G.C., G.D., G.M., A.B., and B.A.W. wrote the paper. All authors reviewed the paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

B.A.W. received funding from the Dutch Cancer Society grants KWF-4874 and KWF-11038, Brain Tumor Charity Grant 488097, Innovation Exchange Amsterdam APCA grant PoC-2017, and the AI-IMPACT and the Toxicity Atlas grants from Health ∼ Holland.

■ ABBREVIATIONS

ABC, ATP-binding cassette transporter; ADME, absorption, distribution, metabolism, and excretion; AUC, area under the curve; BBB, blood−brain barrier; ChEMBL, Cheminformatics databases of the European Molecular Biology Laboratory; CNS, central nervous system; ECFP4, extended-connectivity fingerprints 4; ER, efflux ratio; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; IVIVE, in vitro-to-in vivo; Log*D*, log of partition of a chemical compound between the lipid and aqueous phases;; Log*P*, partition coefficient of a molecule between an aqueous and lipophilic phases; Log*S*, water solubility; MACCS, molecular ACCess System; MW, molecular weight; N.S, nonsignificant; NPV, negative predictive value; PAMPA, parallel artificial membrane permeability assay; *P*-gp, *P*-glycoprotein; PPV, positive predicted value; RBs, rotatable bonds; ROC, responder operator curve; SLCs, solute carriers; SMARTS, list of substructure patterns; TPSA, topological polar surface area; *t*SNE, *t*-distributed stochastic neighbor embedding; XGBoost, eXtreme gradient boosting

■ REFERENCES

(1) Carpenter, T. S.; Kirshner, D.; Lau, E.; Wong, S.; Nilmeier, J.; Lightstone, F. A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophys. J.* **2014**, *107*, 630−641.

(2) Di, L.; Rong, H.; Feng, B. Demystifying Brain Penetration in Central Nervous System Drug Discovery. *J. Med. Chem.* **2012**, *56*, 2−12.
(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) Heffron, T. P. Small Molecule Kinase Inhibitors for the Treatment of Brain Cancer. *J. Med. Chem.* **2016**, *59*, 10030−10066.
(5) Pardridge, W. M. The blood-brain barrier: bottleneck in brain drug development. *NeuroRx* **2005**, *2*, 3−14.
(6) Heffron, T. P.; Salphati, L.; Alicke, B.; Cheong, J.; Dotson, J.; Edgar, K.; Goldsmith, R.; Gould, S. E.; Lee, L. B.; Lesnick, J. D.; et al. The Design and Identification of Brain Penetrant Inhibitors of Phosphoinositide 3-Kinase α. *J. Med. Chem.* **2012**, *55*, 8007−8020.
(7) 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.
(8) Engelhardt, B.; Sorokin, L. The blood−brain and the blood−cerebrospinal fluid barriers: function and dysfunction. *Semin. Immunopathol.* **2009**, *31*, 497−511.
(9) Di, L.; Artursson, P.; Avdeef, A.; Ecker, G. F.; Faller, B.; Fischer, H.; Houston, J. B.; Kansy, M.; Kerns, E. H.; Krämer, S. D.; et al. Evidence-based approach to assess passive diffusion and carrier-mediated drug transport. *Drug Discov. Today* **2012**, *17*, 905−912.
(10) Joó, F. The blood-brain barrier in vitro: The second decade. *Neurochem. Int.* **1993**, *23*, 499−521.
(11) Choi, T. B.; Pardridge, W. M. Phenylalanine transport at the human blood-brain barrier. Studies with isolated human brain capillaries. *J. Biol. Chem.* **1986**, *261*, 6536−6541.
(12) Hargreaves, K. M.; Pardridge, W. M. Neutral amino acid transport at the human blood-brain barrier. *J. Biol. Chem.* **1988**, *263*, 19392−19397.
(13) Matsson, P.; Fenu, L. A.; Lundquist, P.; Wiśniewski, J. R.; Kansy, M.; Artursson, P. Quantifying the impact of transporters on cellular drug permeability. *Trends Pharmacol. Sci.* **2015**, *36*, 255−262.
(14) Schaller, L.; Lauschke, V. M. The genetic landscape of the human solute carrier (SLC) transporter superfamily. *Hum. Genet.* **2019**, *138*, 1359−1377.
(15) Nałęcz, K. A. Solute Carriers in the Blood−Brain Barier: Safety in Abundance. *Neurochem. Res.* **2016**, *42*, 795−809.
(16) César-Razquin, A.; Snijder, B.; Frappier-Brinton, T.; Isserlin, R.; Gyimesi, G.; Bai, X.; Reithmeier, R.; Hepworth, D.; Hediger, M.; Edwards, A.; et al. A Call for Systematic Research on Solute Carriers. *Cell* **2015**, *162*, 478−487.
(17) Garibsingh, R.-A. A.; Schlessinger, A. Advances and Challenges in Rational Drug Design for SLCs. *Trends Pharmacol. Sci.* **2019**, *40*, 790−800.
(18) Nyquist, M. D.; Prasad, B.; Mostaghel, E. A. Harnessing Solute Carrier Transporters for Precision Oncology. *Molecules* **2017**, *22*, 539.
(19) Roth, M.; Obaidat, A.; Hagenbuch, B. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. *Br. J. Pharmacol.* **2012**, *165*, 1260−1287.
(20) Moitra, K.; Dean, M. Evolution of ABC transporters by gene duplication and their role in human disease. *Biol. Chem.* **2011**, *392*, DOI: 10.1515/bc.2011.006.
(21) Miller, D. S. Regulation of P-glycoprotein and other ABC drug transporters at the blood-brain barrier. *Trends Pharmacol. Sci.* **2010**, *31*, 246−254.
(22) Bauer, B.; Hartz, A. M. S.; Fricker, G.; Miller, D. S. Modulation of p-Glycoprotein Transport Function at the Blood-Brain Barrier. *Exp. Biol. Med.* **2005**, *230*, 118−127.
(23) Schinkel, A. H. P. P-Glycoprotein, a gatekeeper in the blood−brain barrier. *Adv. Drug Deliv. Rev.* **1999**, *36*, 179−194.
(24) Vodenkova, S.; Buchler, T.; Cervena, K.; Veskrnova, V.; Vodicka, P.; Vymetalkova, V. 5-fluorouracil and other fluoropyrimidines in colorectal cancer: Past, present and future. *Pharmacol. & Ther.* **2020**, *206*, 107447.
(25) Lanevskij, K.; Dapkunas, J.; Juska, L.; Japertas, P.; Didziapetris, R. QSAR Analysis of Blood−Brain Distribution: The Influence of Plasma and Brain Tissue Binding. *J. Pharm. Sci.* **2011**, *100*, 2147−2160.
(26) Chen, H.; Winiwarter, S.; Engkvist, O. In SilicoTools for Predicting Brain Exposure of Drugs. *Blood-Brain Barrier in Drug Discovery* **2015**, 167−187.
(27) Hou, T.; Wang, J. Structure − ADME relationship: still a long way to go? *Expert Opin. Drug Metab. & Toxicol.* **2008**, *4*, 759−770.
(28) Merlot, C. Computational toxicology—a tool for early safety evaluation. *Drug Discov. Today* **2010**, *15*, 16−22.
(29) Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099−3105.
(30) Miljković, F.; Martinsson, A.; Obrezanova, O.; Williamson, B.; Johnson, M.; Sykes, A.; Bender, A.; Greene, N. Machine Learning Models for Human *In Vivo* Pharmacokinetic Parameters with In-House Validation. *Mol. Pharm.* **2021**, *18*, 4520−4530.
(31) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(32) Saxena, D.; Sharma, A.; Siddiqui, M. H.; Kumar, R. Development of Machine Learning Based Blood-brain Barrier Permeability Prediction Models Using Physicochemical Properties, MACCS and Substructure Fingerprints. *Curr. Bioinform.* **2021**, *16*, 855−864.
(33) 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.
(34) 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.
(35) Castillo-Garit, J. A.; Casanola-Martin, G. M.; Le-Thi-Thu, H.; Pham-The, H.; Barigye, S. J. A Simple Method to Predict Blood-Brain Barrier Permeability of Drug- Like Compounds Using Classification Trees. *Med. Chem.* **2017**, *13*, 664−669.
(36) Muehlbacher, M.; Spitzer, G. M.; Liedl, K. R.; Kornhuber, J. Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. *J. Comput. Aided. Mol. Des.* **2011**, *25*, 1095−1106.
(37) Chen, T. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
(38) Feher, M.; Sourial, E.; Schmidt, J. M. A simple model for the prediction of blood−brain partitioning. *Int. J. Pharm.* **2000**, *201*, 239−247.
(39) Gupta, S.; Basant, N.; Singh, K. P. Qualitative and quantitative structure−activity relationship modelling for predicting blood-brain barrier permeability of structurally diverse chemicals. *SAR QSAR Environ. Res.* **2015**, *26*, 95−124.
(40) Shityakov, S.; Neuhaus, W.; Dandekar, T.; Förster, C. Analysing molecular polar surface descriptors to predict blood-brain barrier permeation. *Int. J. Comput. Biol. Drug Des.* **2013**, *6*, 146.
(41) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Chretien, J. R.; Raevsky, O. A. Estimation of Blood-Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H-Bonding Descriptors. *J. Drug Target.* **1998**, *6*, 151−165.
(42) Rankovic, Z. CNS Drug Design: Balancing Physicochemical Properties for Optimal Brain Exposure. *J. Med. Chem.* **2015**, *58*, 2584−2608.
(43) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold diversity of natural products: inspiration for combinatorial library design. *Nat. Prod. Rep.* **2008**, *25*, 892.
(44) Li, D.; Chen, L.; Li, Y.; Tian, S.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery. 13. Development of *in Silico* Prediction

Models for P-Glycoprotein Substrates. *Mol. Pharm.* **2014**, *11*, 716−726.
(45) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817−834.
(46) Varma, M. V. S.; Sateesh, K.; Panchagnula, R. Functional Role of P-Glycoprotein in Limiting Intestinal Absorption of Drugs: Contribution of Passive Permeability to P-Glycoprotein Mediated Efflux Transport. *Mol. Pharm.* **2004**, *2*, 12−21.
(47) Gross, G.. *Chapter 27 - Strategies for Enhancing Oral Bioavailability and Brain Penetration*. in Wermuth, C. G., Aldous, D., Raboisson, P.; Rognan, D. B. T.-T. P., Eds; Academic Press, 2015 Fourth E pp 631−655
(48) Matsson, P.; Bergström, C. A. S. Computational modeling to predict the functions and impact of drug transporters. *silico Pharmacol* **2015**, *3*, 8.
(49) Yamauchi, S.; Sugano, K. Permeation characteristics of tetracyclines in parallel artificial membrane permeation assay. *ADMET DMPK* **2019**, *7*, 151−160.
(50) van de Waterbeemd, H.; Smith, D. A. Relations of Molecular Properties with Drug Disposition: The Cases of Gastrointestinal Absorption and Brain Penetration. *Pharmacokinetic Optimization in Drug Research* **2007**, 51−64.
(51) Akanuma, S.; Uchida, Y.; Ohtsuki, S.; Kamiie, J. i.; Tachikawa, M.; Terasaki, T.; Hosoya, K. i. Molecular-weight-dependent, Anionic-substrate-preferential Transport of β-Lactam Antibiotics via Multidrug Resistance-associated Protein 4. *Drug Metab. Pharmacokinet.* **2011**, *26*, 602−611.
(52) Nau, R.; Sörgel, F.; Eiffert, H. Penetration of drugs through the blood-cerebrospinal fluid/blood-brain barrier for treatment of central nervous system infections. *Clin. Microbiol. Rev.* **2010**, *23*, 858−883.
(53) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034−1041.
(54) Adenot, M.; Lahana, R. Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *35*, 239−248.
(55) Glem, R. C.; et al. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199−204.
(56) Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine tools: An online service for analyzing and clustering small molecules. *Nucleic Acids Res.* **2011**, *39*, 486−491.
(57) Ghose, A. K.; Herbertz, T.; Hudkins, R. L.; Dorsey, B. D.; Mallamo, J. P. Knowledge-Based, Central Nervous System (CNS) Lead Selection and Lead Optimization for CNS Drug Discovery. *ACS Chem. Neurosci.* **2012**, *3*, 50−68.
(58) Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. J. Predicting blood-brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis. *Pharm. Res.* **2002**, *19*, 1611−1621.
(59) Ly, K. I.; Wen, P. Y. Clinical Relevance of Steroid Use in Neuro-Oncology. *Curr. Neurol. Neurosci. Rep.* **2017**, *17*, 5.
(60) Reardon, D. A.; Desjardins, A.; Vredenburgh, J. J.; Gururangan, S.; Friedman, A. H.; Herndon, J. E.; Marcello, J.; Norfleet, J. A.; McLendon, R. E.; Sampson, J. H.; et al. Phase 2 trial of erlotinib plus sirolimus in adults with recurrent glioblastoma. *J. Neurooncol.* **2010**, *96*, 219−230.
(61) Peng, J.; Hu, Q.; Gu, C.; Liu, B.; Jin, F.; Yuan, J.; Feng, J.; Zhang, L.; Lan, J.; Dong, Q.; et al. Discovery of potent and orally bioavailable inhibitors of Human Uric Acid Transporter 1 (hURAT1) and binding mode prediction using homology model. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 277−282.
(62) Lin, Y.; Chen, X.; Ding, H.; Ye, P.; Gu, J.; Wang, X.; Jiang, Z.; Li, D.; Wang, Z.; Long, W.; et al. Efficacy and safety of a selective URAT1 inhibitor SHR4640 in Chinese subjects with hyperuricaemia: a randomized controlled phase II study. *Rheumatology (Oxford)* **2021**, *60*, 5089−5097.
(63) Mohs, R. C.; Greig, N. H. Drug discovery and development: Role of basic biological research. *Alzheimer's Dement. (New York, N. Y.)* **2017**, *3*, 651−657.
(64) Cocucci, E.; Kim, J. Y.; Bai, Y.; Pabla, N. Role of Passive Diffusion, Transporters, and Membrane Trafficking-Mediated Processes in Cellular Drug Transport. *Clin. Pharmacol. & Ther.* **2016**, *101*, 121−129.
(65) Di, L.; Rong, H.; Feng, B. Demystifying brain penetration in central nervous system drug discovery. *J. Med. Chem.* **2013**, *56*, 2−12.
(66) Sugano, K.; Kansy, M.; Artursson, P.; Avdeef, A.; Bendels, S.; Di, L.; Ecker, G. F.; Faller, B.; Fischer, H.; Gerebtzoff, G.; et al. Coexistence of passive and carrier-mediated processes in drug transport. *Nat. Rev. Drug Discovery* **2010**, *9*, 597−614.
(67) Kell, D. B.; Dobson, P. D.; Oliver, S. G. Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. *Drug Discov. Today* **2011**, *16*, 704−714.
(68) Kumar, S.; Das, A. Effect of acceptor heteroatoms on π-hydrogen bonding interactions: A study of indole···thiophene heterodimer in a supersonic jet. *J. Chem. Phys.* **2012**, *137*, 094309.
(69) Heffron, T. P. Small Molecule Kinase Inhibitors for the Treatment of Brain Cancer. *J. Med. Chem.* **2016**, *59*, 10030−10066.
(70) Karthikeyan, M.; Glen, R. C.; Bender, A. General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks. *J. Chem. Inf. Model.* **2005**, *36*, 581−590.
(71) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian Approach to *in Silico* Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686−1697.
(72) Wang, W.; Kim, M. T.; Sedykh, A.; Zhu, H. Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* **2015**, *32*, 3055−3065.
(73) 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.
(74) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.; Mosquera, J.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(75) Mak, L.; Marcus, D.; Howlett, A.; Yarova, G.; Duchateau, G.; Klaffke, W.; Bender, A.; Glen, R. C. Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *J. Cheminform.* **2015**, *7*, 31.
(76) Miller, D. S. ABC Transporter Regulation by Signaling at the Blood−Brain Barrier. *Pharmacology of the Blood Brain Barrier Targeting CNS Disorders* **2014**, 1−24.
(77) Miller, D. S. Regulation of ABC Transporters Blood−Brain Barrier. *ABC Transporters and Cancer* **2015**, 43−70.
(78) Miller, D.; Cannon, R. Signaling Pathways that Regulate Basal ABC Transporter Activity at the Blood- Brain Barrier. *Curr. Pharm. Des.* **2014**, *20*, 1463−1471.
(79) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J. Med. Chem.* **1998**, *41*, 1007−1010.
(80) Wohnsland, F.; Faller, B. High-Throughput Permeability pH Profile and High-Throughput Alkane/Water log *P* with Artificial Membranes. *J. Med. Chem.* **2001**, *44*, 923−930.
(81) Di, L.; Kerns, E. H.; Fan, K.; McConnell, O. J.; Carter, G. T. High throughput artificial membrane permeability assay for blood−brain barrier. *Eur. J. Med. Chem.* **2003**, *38*, 223−232.
(82) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.
(83) Bento, A. P.; Hersey, A.; Félix, E.; Landrum, G.; Gaulton, A.; Atkinson, F.; Bellis, L. J.; De Veij, M.; Leach, A. R. An open source chemical structure curation pipeline using RDKit. *J. Cheminform.* **2020**, *12*, 51.

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267

(84) Guha, R. Chemical informatics functionality in R. *J. Stat. Softw.* **2007**, *18*, 1−16.
(85) R Core Team. R. *A Language and Environment for Statistical Computing*: Vienna,Austria, 2020.
(86) Landrum, G. *RDKit: Open-Source Cheminformatics*, 2019. (Accessed 10.
(87) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(88) Kuhn, M. *Classification and Regression Training*, 2022.

https://doi.org/10.1021/acs.jmedchem.2c01824
J. Med. Chem. 2023, 66, 7253−7267