A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors

Fanwang Meng, Yang Xi, Jinfeng Huang & Paul W. Ayers

The highly-selective blood-brain barrier (BBB) prevents neurotoxic substances in blood from crossing into the extracellular fluid of the central nervous system (CNS). As such, the BBB has a close relationship with CNS disease development and treatment, so predicting whether a substance crosses the BBB is a key task in lead discovery for CNS drugs. Machine learning (ML) is a promising strategy for predicting the BBB permeability, but existing studies have been limited by small datasets with limited chemical diversity. To mitigate this issue, we present a large benchmark dataset, B3DB, compiled from 50 published resources and categorized based on experimental uncertainty. A subset of the molecules in B3DB has numerical log BB values (1058 compounds), while the whole dataset has categorical (BBB+ or BBB−) BBB permeability labels (7807). The dataset is freely available at https://github.com/theochem/B3DB and https://doi.org/10.6084/m9.figshare.15634230.v3 (version 3). We also provide some physicochemical properties of the molecules. By analyzing these properties, we can demonstrate some physicochemical similarities and differences between BBB+ and BBB− compounds.

Background & Summary

The blood-brain barrier (BBB) denotes a regulatory and protective mechanism of microvasculature in the central nervous system (CNS) that is central to regulating the homeostasis of the CNS and protecting the CNS from toxins, pathogens, and inflammations. However, it is estimated that 98% of small molecule drugs are not BBB permeable. Therefore, predicting BBB permeability for small molecules is a vital but challenging task in drug discovery and development. However, existing computational models for a molecule's BBB permeability are inadequate. In particular, they are restricted by the limited size and chemical diversity of existing sets of training data. Moreover, although many different machine-learning (ML) models for predicting BBB permeability have been proposed, these models are not directly comparable because they use widely varying training data, ranging from as few as 45 molecules to as many as 7236 molecules. The purpose of this paper is to curate an accessible, clean, well-documented, and reasonably comprehensive dataset of BBB permeability data and present it in a way that is convenient for those building new BBB predictive models. While our database, B3DB, is not the first attempt to curate data from the literature to construct a molecular BBB database, B3DB contains more molecules, and categorizes the molecules based on experimental uncertainty. Both features are very helpful when developing and validating ML models for BBB.

There are two types of data for BBB, numerical and categorical data. Numerical data is usually reported as log BB, the logarithm of brain-plasma concentration ratio,

\[ \log BB = \log \frac{C_{\text{brain}}}{C_{\text{blood}}} \]  

Categorical data simply labels whether a compound is BBB permeable (BBB+) or not (BBB−).

Among existing studies of BBB permeability, we mention Zhuang et al., who built a ML model with resampling using a binary dataset of 2358 molecules. Similarly, Zhao et al. compiled a dataset of 1336 BBB compounds.
crossing drugs (BBB+) and 360 BBB non-crossing drugs (BBB−). To our knowledge, the largest dataset previously reported in the literature was used in developing the lightBBB model, which uses the Light Gradient Boosting Machine (LightGBM) algorithm to build a predictive model. The lightBBB model’s database included 7162 entries. (These entries include duplicates (multiple entries with the same International Chemical Identifier (InChI)) and molecules that could not be recognized by RDKit, so in the end there are only 4491 unique valid molecules). We curate data from these three efforts, and 47 other smaller efforts, in B3DB. Unlike many previous efforts, B3DB includes many (1058) molecules with numeric log BB values. The largest previous dataset we know was the data source for lightBBB, which has log BB values for 696 unique valid molecules.

Here, we present a new Blood-Brain Barrier Database, B3DB, which is intended to provide a benchmark dataset for modelling BBB permeability of small molecules. The original data was collected from 50 peer-reviewed publications or open access datasets. As described in the next section, we processed and cleaned the data, then categorized it based on its reliability. By categorizing the data in this way, users can choose whether they want to focus on the smaller subsets with the highest reliability, or prefer to consider larger datasets with slightly lower reliability. We hope that our meticulous methods of preparing and sorting the data may be of interest those who wish to curate databases for other, similar, properties.

B3DB includes both numerical data (1058 log BB values) and categorical data (4956 BBB+ and 2851 BBB−). Here is summary of key features of B3DB dataset. (1) This is the largest BBB database we know, both for categorical labels and log BB numerical values. (2) Because the chirality of molecules plays an important role in BBB permeability14,15, isomeric SMILES is to used to incorporate chiral specifications of molecules. (3) Because some molecules have been measured multiple times, using different experimental methods and under different conditions, we divide the value into groups based on the quantity of experimental data and the similarity between reported values, so that users of B3DB can easily select subsets of the data with varying degrees of reliability. (4) B3DB is extended with molecular descriptors computed with mordred16, so that it can be used out-of-the-box for building BBB predictive models.

Methods

The next three sections describe how raw data was collected from various sources, cleaned, and curated. We then describe how the dataset was extended with chemical descriptors (beyond the reference BBB value). This workflow is summarized in Fig. 1. For consistency and reproducing purposes, all the data processing were performed in a Python 3.7.9 environment created with Conda in CentOS Linux release 7.9.2009 which include Numpy 1.19.2, mordred 1.1.1, PyTDC 0.1.5. ALOGPS version 2.1 is also used for calculating octanol/water partition coefficient log P.

Data collecting. All the data was collected from the literature and open source databases. The dataset size, main available information, and data types are listed in Table 1. For each data source, a standard Excel workbook is formatted for further processing. If the original data is in a portable document format (PDF), it is converted to a pandas DataFrame and then stored in XLSX format with tabula-py18. For files in DOCX or DOC extension, as well as CSV, TXT and other Excel compatible formats, they are converted to Excel XLSX format directly, using Microsoft Office. We performed several automated consistency checks (e.g., numerical data should be reported as floating-point numbers) and manually verified a subset of the data to ensure that the data was faithfully transferred to XLSX format. In total, 33825 raw data records were collected.

The 50 datasets have various formats and include a wide range of information, so we constructed a template that contained only the most essential data, compound name, simplified molecular-input line-entry system (SMILES) string, PubChem compound identifier (CID), log BB, BBB+/BBB− (whether a compound is BBB permeable or not), the IUPAC International Chemical Identifier (InChI), the threshold value used to determine categorical type of a compound, and the literature source for that data value.

Data cleaning. In the data cleaning stage, an initial molecule specification (a SMILES string, PubChem CID, and/or compound name) is input; the output is also a SMILES string, but with transcription and typographical
errors fixed, and with salts/solvents removed. In addition, molecules containing heavy metal atoms are removed from the database. A followed up standardization of molecular reorientation is performed which include updating valences, kekulizing and normalizing molecules, and neutralizing molecular charges. The basic procedure is shown in Fig. 2(a).

Table 1. Data source and the available corresponding information. *Data accessed with PyTDC 0.1.5 as of Jan 25, 2021.

| ID | Data Source Size | Information Available | Data Type | Reference |
|----|------------------|-----------------------|-----------|-----------|
| R1 | 2053             | name, smiles         | categorical data | 20        |
| R2 | 1210             | name, smiles         | categorical data, numerical data | 22        |
| R3 | 328              | name, smiles         | numerical data | 33        |
| R4 | 189              | CAS, name, smiles    | numerical data | 36        |
| R5 | 108              | name, smiles         | numerical data | 37        |
| R6 | 1692             | name, smiles         | categorical data | 38        |
| R7 | 224              | name                 | categorical data | 29        |
| R8 | 439              | smiles, CID          | numerical data | 25        |
| R9 | 415              | name, smiles         | categorical data | 30        |
| R10| 462              | name, CID            | categorical data | 40        |
| R11| 151              | name, logBB          | numerical data | 41        |
| R12| 182              | name, smiles         | numerical data | 42        |
| R13| 2321             | smiles               | categorical data | 11        |
| R14| 942              | name, smiles         | categorical data | 43        |
| R15| 390              | name                 | categorical data | 44        |
| R16| 374              | name, CID            | categorical data | 30        |
| R17| 55               | name                 | numerical data | 45        |
| R18| 332              | name, smiles         | numerical data | 27        |
| R19| 1990             | name, smiles         | categorical data | 13        |
| R20| 139              | name                 | numerical data | 46        |
| R21| 362              | name, smiles, CID    | numerical data | 47        |
| R22| 27               | name                 | numerical data | 48        |
| R23| 1090             | name, smiles         | categorical data | 49        |
| R24| 1866             | smiles               | categorical data | 50        |
| R25| 581              | name, smiles         | numerical data | 26        |
| R26| 448              | CAS, name, smiles    | categorical data, numerical data | 51        |
| R27| 7236             | smiles               | categorical data, numerical data | 11        |
| R28| 415              | name, smiles         | categorical data | 31        |
| R29| 181              | name                 | categorical data | 32        |
| R30| 3620             | name, smiles         | categorical data | 53        |
| R31| 12               | name                 | numerical data | 34        |
| R32| 26               | name                 | numerical data | 35        |
| R33| 26               | name                 | numerical data | 36        |
| R34| 153              | name                 | numerical data | 37        |
| R35| 145              | smiles               | numerical data | 58        |
| R36| 525              | name, smiles         | categorical data | 59        |
| R37| 111              | name, smiles         | categorical data | 60        |
| R38| 291              | name, smiles         | numerical data | 61        |
| R39| 122              | name                 | numerical data | 62        |
| R40| 405              | name                 | numerical data | 63        |
| R41| 296              | smiles               | numerical data | 64        |
| R42| 45               | smiles               | numerical data | 1         |
| R43| 328              | name, smiles         | numerical data | 65        |
| R44| 89               | name                 | numerical data | 66        |
| R45| 8                | smiles               | numerical data | 67        |
| R46| 483              | smiles               | numerical data | 68        |
| R47| 529              | name                 | numerical data | 69        |
| R48| 115              | smiles               | numerical data | 70        |
| R49| 181              | name, smiles         | numerical data | 71        |
| R50| 113              | name, smiles         | categorical data, numerical data | 72        |
The first step is to fix invalid SMILES strings. For example, white spaces and line breaks in SMILES were removed. Some other issues (e.g., where a dash was used in lieu of a negative sign for the molecular charge) were manually remedied. Our data is drawn from 50 distinct sources, and a full molecule specification is not always provided. For example, some sources list only the compound names (and not the SMILES strings or PubChem CIDs); other sources list only PubChem CIDs. In these cases, PubChemPy was used to access the PubChem database to retrieve information about missing compound names, SMILES strings and PubChem CIDs. When only the compound name was available, there can be multiple PubChem instances. If this were to happen, the first PubChem instance is selected and a note is added to the database flagging the potential ambiguity. Fortunately this does not seem to occur in this specific database. There are also a few molecules for which only molecular structures, and not SMILES or compound names, are provided. In these cases we built the molecules manually and searched for the Pubchem CID and SMILES string with the PubChem web interface. All the SMILES strings were loaded into RdKit (version 2019.03.4) to build molecule objects. If the object is None, the SMILES is neutralized. Molecules containing metal atoms or heavy atoms with atomic number greater than 20 (e.g., Zinc, Bromine, Krypton, Iodine, and Xenon) were removed. Molecules with more than 7 boron atoms and neutralize the charge. Molecules containing metal atoms or heavy atom with atomic number greater than 20 (e.g., Zinc, Bromine, Krypton, Iodine, and Xenon) were removed. Molecules with more than 7 boron atoms are also excluded due to problems of depicting borane compounds. Implicit valence and ring information were recomputed followed by kekulizing, normalization of molecules and molecular charges were neutralized. These revisions change the molecular structure, so the Pubchem CIDs were updated from the revised SMILES strings.

**Data curation.** The curation procedures for numerical and categorical data are summarized in Fig. 3. To curate the data, a unique chemical identifier is required. Although InChI is unique in principle, it cannot resolve tautomeric forms, which is a common source of ambiguity and error in chemical structure representation. Therefore, we examined the unique InChI generated with RdKit and the isomeric SMILES (and canonical SMILES where isomeric SMILES is unavailable). The number of unique SMILES is greater than the number of unique InChI values, but the redundancy is merely because each SMILES represents a specific resonance structure.

Data curation.  The curation procedures for numerical and categorical data are summarized in Fig. 3. To curate the data, a unique chemical identifier is required. Although InChI is unique in principle, it cannot resolve tautomeric forms, which is a common source of ambiguity and error in chemical structure representation. Therefore, we examined the unique InChI generated with RdKit and the isomeric SMILES (and canonical SMILES where isomeric SMILES is unavailable). The number of unique SMILES is greater than the number of unique InChI values, but the redundancy is merely because each SMILES represents a specific resonance structure.

Curation of numerical data. To curate the 8841 numerical BBB data values, log BB values for each molecule were merged into a list. The 20 instances with log BB ≤ −9 were regarded as outliers because, based on the distribution of log BB values, they seemed suspicious. Next, we identified molecules where there are multiple reported log BB values and eliminated those molecules from the database if the reported values differed significantly. Specifically, we eliminated 16 molecules where max (log BB) − min(log BB) > 1. The values that remain after curation are merged into 1065 molecular records. The molecular records are augmented, as necessary, to ensure that they are complete, including compound name, IUPAC name, isomeric (canonical) SMILES, etc.

Here is the detailed curation procedure for numeric data.

1. **Group A** (243 molecules). Molecules with only one unique log BB value.
2. **Group B** (663 molecules). Molecules with more than one log BB value, but all the the reported values differ

![Fig. 2](https://www.nature.com/scientificdata/)

**Fig. 2** Molecule representation cleaning and technical validation. (a) Flowchart of cleaning SMILES string representation of molecules. (b) Technical validation of molecular representation.
by less than 5% from the mean value. In these cases, the mean value is used as the log \( BB \) value for the molecule.

3. **Group C** (3 molecules). Other molecules with two distinct log \( BB \) values. The (weighted) mean value is used as the curated value for group C (just as for group B).

4. **Group D** (149 molecules). Other molecules with more than two distinct values; whichever value occurs with greatest frequency is used. In three cases, two distinct values were reported with maximum frequency; we discarded those molecules from the dataset.

The 7 molecules which failed to be categorized as group A, B, C or D, they are discarded. The final dataset therefore contains 1058 molecules; for most of these molecules (815 molecules) multiple, mutually consistent, values of log \( BB \) are reported in the literature.

**Curation of categorical data.** The 33689 data values were divided into two categories, numerical data and (binary) categorical data.

1. **Group A** (1058 molecules). Molecules with numerical data. Several threshold values for log \( BB \) have been used to determine if a molecule is BBB permeable or not, including \( 0^{25,26}, 0.127^{13,28–31}, (−2, 1)^{32} \). The value of \( −1 \) is chosen as the threshold value to define if a compound is BBB+ or BBB− since this is the mostly widely used threshold and maximizes the ease of comparison with other studies.

2. **Group B** (3621 molecules). Molecules from sources that use log \( BB = −1 \) as the threshold value, and where all sources agree on the categorical label. The unambiguous label is used.

3. **Group C** (3077 molecules). Molecules where all sources agree on the categorical label, but the sources that do not report their threshold value.

4. **Group D** (51 molecules). Molecules with two different BBB permeability labels. The most prevalent label is used. In the 45 cases where the two labels occurred with equal frequency, the molecule was discarded.

The 7807 remaining molecular records are augmented to ensure that they are complete, including compound name, IUPAC name, isomeric (canonical) SMILES, etc..

**Data extension with chemical descriptors.** To better facilitate building BBB predictive models, the curated datasets were extended with chemical descriptors. Then 1613 chemical descriptors were calculated with mordred version 1.1.16. The purpose of providing this extended data is to facilitate easy use of the B3DB, without requiring precomputation of cheminformatics descriptors.

**Data Records**

There are two datasets provided in this study, one with numeric log \( BB \) values (1058 molecules) and the other with categorical labels (7807 molecules with 4956 BBB+ and 2851 BBB−). B3DB data is stored in the comma-separated values (CSV) format and contains SMILES representations, compound name, IUPAC name, log \( BB \) value, threshold, BBB+/BBB− and the corresponding references along with 1613 molecular descriptors. This is summarized in Table 2. The data are openly accessible at GitHub (https://github.com/theochem/B3DB) as well as figshare platform33.
Technical Validation

Validation of molecular representations. All the molecules are in canonical SMILES format and, if available from PubChem, also isomeric SMILES. We then attempt to load each SMILES string into OEChem Toolkit\(^1\) as an OEGraphMol object; if this is successful then this SMILES is regarded as valid. (See Fig. 2(b)).

Analysis of curated datasets. The BBB data comes from 50 sources, and was acquired in different laboratories, under different conditions, and using different protocols. To characterize the experimental uncertainty, we examine the agreement between reported values, Fig. 4. For 92.82% of the numerical data, there at most two unique log BB values are reported as shown in Fig. 4(a,c). Similarly, for 99.34% of the molecules, only a single categorical label is reported (Fig. 4(d)); this is true even though the same molecule may appear in as many as 23 distinct sources (Fig. 4(b)). More detailed data can be found in Tables 3–6.

Figure 5 reveals some features of the B3DB dataset. Presuming that the molecules in the dataset are relatively representative of (bio)organic molecules in general, the log BB for most of organic compounds lie within the interval \([-2, 2]\) (see Fig. 5(a)). The distribution of log BB values indicates that the numerical dataset is relatively balanced, though skewed towards BBB\(^+\) compounds.

Lipinski's Rule of 5 \(\text{https://www.sciencedirect.com/science/article/abs/pii/S0169409X00001290?via%3Dihub}\) is a simple rule-of-thumb for evaluating a molecule's drug-likeness. Specifically, Lipinski's Rule of 5 states that good absorption or permeation is more likely if a molecule has less than: 5 hydrogen-bond donors, 10

---

| Column Header | Description                                                                 | Data Type |
|---------------|-----------------------------------------------------------------------------|-----------|
| compound name | Generic name of compound                                                    | string    |
| IUPAC name    | Name of compound following the IUPAC nomenclature naming scheme             | string    |
| SMILES        | SMILES representation of compound, isomeric SMILES if available             | string    |
| CID           | PubChem compound identifier                                                 | string    |
| log BB        | log BB value of compound                                                    | float     |
| BBB+/BBB−     | Categorical labels to indicate if compound is BBB permeable (BBB+) or not (BBB−) | string    |
| InChI         | The IUPAC International Chemical Identifier of compound                      | string    |
| threshold     | Threshold value used to determine BBB permeability label                    | float     |
| reference     | Data sources                                                                | string    |
| group         | Group classification                                                        | string    |
| comment       | Complementary information                                                   | string    |

Table 2. List of information in the curated datasets. The BBB+/BBB− and threshold columns are only available for categorical dataset. The 1613 2D chemical descriptors are not listed in this table.
hydrogen-bond acceptors, 500 Dalton molecular weight, and a predicted log $P$ value less than 5. It is observed that the molecule weight of most BBB$^+$ compounds (93.10%) is less than 500 Dalton. In contrast, there are many molecules with molecular weight greater than 500 Dalton (31.22%) that are BBB$^-$ compounds. Nonetheless, aside from the a long tail of heavy BBB$^-$ compounds, the distribution of molecular weights for BBB$^+$ and BBB$^-$ molecules is not dissimilar (see Fig. 5(b,f)). 98.8% of BBB$^+$ compounds and 23.4% of BBB$^-$ compounds have fewer than 5 hydrogen-bond donors; 97.6% of BBB$^+$ compounds and 66.0% of BBB$^-$ compounds have fewer than 10 hydrogen-bond acceptors. This supports the idea that hydrophilic compounds find it difficult to cross

| Group | A | B | C | D |
|-------|---|---|---|---|
| 1     | 243 | 0 | 0 | 0 |
| 2     | 0  | 9 | 3 | 8 |
| 3     | 0  | 2 | 0 | 29|
| 4     | 0  | 0 | 0 | 21|
| 5     | 0  | 0 | 0 | 8 |
| 6     | 0  | 0 | 0 | 7 |

Table 3. Occurrences of source log $BB$ values for different groups in numerical dataset.

| Group | A | B | C | D |
|-------|---|---|---|---|
| 1     | 652 | 0 | 0 | 0 |
| 2     | 9  | 3 | 84| 5 |
| 3     | 2  | 0 | 29| 1 |
| 4     | 0  | 0 | 21| 2 |
| 5     | 0  | 0 | 8 | 2 |
| 6     | 0  | 0 | 7 | 3 |

Table 4. Occurrences of unique source log $BB$ values for different groups in numerical dataset.
the BBB, but this is not a hard-and-fast rule: there are BBB+ compounds that violate Lipinski’s rule of 5. Finally, the octanol/water partition coefficient log $P$ was estimated using ALOPS version 2.134. There is not much difference in the log $P$ values for BBB+ and BBB− compounds: 93.8% of BBB+ and 95.1% of BBB− compounds

---

**Table 5.** Occurrences of source BBB permeability labels for different groups in categorial dataset.
Table 6. Occurrences of unique source BBB permeability labels for different groups in categorical dataset.

| Group | A | B | C | D |
|-------|---|---|---|---|
| 1     | 1058 | 3621 | 3077 | 0  |
| 2     | 0  | 0  | 0  | 51 |

have log $P < 5$. Taken together, the analysis of the selected physiochemical descriptors suggest that no single parameter can determine the BBB-permeability of a compound. This confirms that predicting BBB permeability computationally is challenging, and emphasizes the value of the B3DB dataset.

Usage Notes
None of the original data sources contain any quantification of uncertainty (e.g., the standard derivation), so it is recommended to incorporate the group categories when using the datasets. If one decides to use a different threshold to determine BBB+ and BBB− for a molecules, log BB can be used directly from the data reported in this study. The 1613 2D chemical descriptors, computed with mordred can facilitate building predictive models. Any further molecular preprocessing can be done with RdKit.

Code availability
The codes used in this study have been deposited to https://github.com/theochem/B3DB and https://doi.org/10.6084/m9.figshare.15634230.v3 (version 3). All the calculation were done with Python 3.7.9 under a virtual environment created with Anaconda on Linux.

Received: 28 July 2021; Accepted: 22 September 2021;
Published online: 29 October 2021

References
1. Risau, W. & Wolburg, H. Development of the blood-brain barrier. Trends in Neurosciences 13, 174–178 (1990).
2. Profaci, C. P., Murji, R. N., Pulido, R. S. & Daneman, R. The blood–brain barrier in health and disease: Important unanswered questions. Journal of Experimental Medicine 217 (2020).
3. Daneman, R. & Prat, A. The blood–brain barrier. Cold Spring Harbor Perspectives in Biology 7, a020412 (2015).
4. Partridge, W. M. Blood–brain barrier delivery. Drug Discovery Today 12, 54–61 (2007).
5. Partridge, W. M. The blood-brain barrier: bottleneck in brain drug development. NeuroRx 2, 3–14 (2005).
6. Harilal, S. et al. Revisiting the blood-brain barrier: A hard nut to crack in the transportation of drug molecules. Brain Research Bulletin (2020).
7. Veszelka, S., Kittel, Á. & Deli, M. A. Tools of modelling blood–brain barrier penetrability. Solubility. Delivery and ADME Problems of Drugs and Drug-Candidates, Bentham Science Publishers, Washington 166–188 (2011).
8. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset’s chemical diversity limits the generalizability of machine learning predictions. Journal of Cheminformatics 11, 69 (2019).
9. Ciura, K. et al. Assessment of blood–brain barrier permeability using micellar electrokinetic chromatography and p_vsa-like descriptors. Microchemical Journal 158, 105236 (2020).
10. Kelder, J., Grotenhuis, P. D., Bayada, D. M., Delbressine, L. P. & Ploemen, J.-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharmaceutical Research 16, 1514–1519 (1999).
11. Shaker, B. et al. Lightbbb: computational prediction model of blood–brain-barrier penetration based on lightgbm. Bioinformatics (2020).
12. Wang, Z. et al. In silico prediction of blood-brain barrier permeability of compounds by machine learning and resampling methods. ChemMedChem 13, 2189–2201 (2018).
13. Zhao, Y. H. et al. Predicting penetration across the blood–brain barrier from simple descriptors and fragmentation schemes. Journal of Chemical Information and Modeling 47, 170–175 (2007).
14. Chang, K. L., Lee, H. N., Yang, S. & Ho, P. C. Influence of drug transporters and stereoselectivity on the brain penetration of pioglitazone as a potential medicine against alzheimer’s disease. Scientific Reports 5, 1–7 (2015).
15. Fong, C. W. Permeability of the blood–brain barrier: molecular mechanism of transport of drugs and physiologically important compounds. The Journal of Membrane Biology 248, 651–669 (2015).
16. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics 10, 1–14 (2018).
17. OEChem, T. Openeye scientific software. Inc., Santa Fe, NM, USA (2020).
18. McKinney, W. Data structures for statistical computing in python. In van der Walt, S. & Millman, J. (eds.) Proceedings of the 9th Python in Science Conference, 51–56 (2010).
19. Ariga, A. tabula-py (2020).
20. Swain, M. Pubchempy (2017).
21. Kim, S. et al. Pubchem in 2021: new data content and improved web interfaces. Nucleic Acids Research 49, D1388–D1395 (2021).
22. Landrum, G. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling (2013).
23. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. An update on pubchem: Nucleic Acids Research 46, W563–W570 (2018).
24. Bento, A. P. et al. An open source chemical structure curation pipeline using rdkit. Journal of Cheminformatics 12, 1–16 (2020).
25. Wang, W., Kim, M. T., Sedykh, A. & Zhu, H. Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in qsar modeling. Pharmaceutical Research 32, 3055–3065 (2015).
26. Brito-Sánchez, Y. et al. Towards better bbb passage prediction using an extensive and curated data set. Molecular Informatics 34, 308–330 (2015).
27. Pisson, F. & Piggott, A. M. Predicting blood–brain barrier permeability of marine-derived kinase inhibitors using ensemble classifiers reveals potential hits for neurodegenerative disorders, Marine Drugs 17, 81 (2019).
28. Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcão, A. O. A bayesian approach to in silico blood-brain barrier penetration modeling. Journal of Chemical Information and Modeling 52, 1686–1697 (2012).
67. Hemmateenejad, B., Miri, R., Safarpour, M. A. & Mehdipour, A. R. Accurate prediction of the blood–brain partitioning of a large set

65. Dichiara, M., Amata, B., Turnaturi, R., Marrazzo, A. & Amata, E. Tuning properties for blood–brain barrier permeation: A statistics-

68. Valencia, C. Y . M.

64. Toropov, A. A., Toropova, A. P ., Beeg, M., Gobbi, M. & Salmona, M. Qsar model for blood-brain barrier permeation.

63. Vilar, S., Chakrabarti, M. & Costanzi, S. Prediction of passive blood–brain partitioning: straightforward and effective classification

62. Shayanfar, A., Soltani, S. & Jouyban, A. Prediction of blood–brain distribution: effect of ionization.

59. Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A. & Tropsha, A. Qsar modeling of the blood–brain barrier permeability for diverse

50. Roy, D., Hinge, V . K. & Kovalenko, A. To pass or not to pass: predicting the blood–brain barrier permeability with the 3d-rism-kh

52. Subramanian, G. & Kitchen, D. B. Computational models to predict blood–brain barrier permeation and cns activity.

51. Chico, L. K., Van Eldik, L. J. & Watterson, D. M. Targeting protein kinases in central nervous system disorders. Nature Reviews Drug

41. Shen, J., Du, Y ., Zhao, Y ., Liu, G. & Yang, Y. In silico prediction of blood–brain partitioning using a chemometric method called
genetic algorithm based variable selection. Qsar & Combinatorial Science 27, 704–717 (2008).

49. Majumdar, S., Basak, S. C., Lungu, C. N., Diudea, M. V . & Grunwald, G. D. Finding needles in a haystack: determining key molecular
descriptors associated with the blood–brain barrier entry of chemical compounds using machine learning. Molecular Informatics 38, 1800164 (2019).

48. 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 Chemical Neuroscience 3, 56–68 (2012).

47. Broccatelli, F., Larregieu, C. A., Cruciani, G., Oprea, T. I. & Benet, L. Z. Improving the prediction of the brain disposition for orally
administered drugs using bddcs. Advanced Drug Delivery Reviews 64, 95–109 (2012).

46. Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. prediction of intestinal absorption. Journal of Pharmaceutical Sciences 88, 807–814 (1999).

45. Gupta, M., Lee, H. J., Barden, C. J. & Weaver, D. F. The Blood-Brain Barrier (BBB) score. Journal of Medicinal Chemistry 62, 9824–9836 (2019).

44. Kortagere, S., Chakmavare, D., Welb, W. J. & Ekins, S. New predictive models for blood–brain barrier permeability of drug-like
molecules. Pharmaceutical Research 25, 1836–1845 (2008).

43. Fu, X.-C., Mang, G.-P., Shan, H.-L., Liang, W.-Q. & Gao, J.-Q. Predicting blood–brain barrier penetration from molecular weight and number of polar atoms. European Journal of Pharmacetics and Biopharmacetics 70, 462–466 (2008).

42. Lanevskij, K., Dapkunas, J., Juska, L., Japertas, P. & Didziapetris, R. Qsar analysis of blood–brain distribution: The influence of plasma and brain tissue binding. Journal of Pharmaceutical Sciences 100, 2147–2160 (2011).

41. Shayanfar, A., Soltani, S. & Jouyban, A. Prediction of blood–brain distribution: effect of ionization.

40. Miao, R., Xia, L.-Y ., Chen, H.-H., Huang, H.-H. & Liang, Y. Improved classification of blood-brain-barrier drugs using deep learning.

39. Majumdar, S., Basak, S. C., Lungu, C. N., Diudea, M. V . & Grunwald, G. D. Finding needles in a haystack: determining key molecular
descriptors associated with the blood–brain barrier entry of chemical compounds using machine learning. Molecular Informatics 38, 1800164 (2019).

38. Adenot, M. & Lahana, R. Blood-brain barrier permeation models: discriminating between potential cns and non-cns drugs including p-glycoprotein substrates. Journal of Chemical Information and Computer Sciences 44, 239–248 (2004).

37. Guerra, A., Páez, J. A. & Campillo, N. E. Artificial neural networks in admet modeling: prediction of blood–brain barrier
permeation. Bulletin of the chemical Society of Japan 83, 25–32 (2010).

36. Mente, S. & Lombardo, F . A recursive-partitioning model for blood–brain barrier permeation. Journal of Computer-Aided Molecular
Design 19, 465–481 (2005).

35. Huang, K.

34. Tetko, I. V . & Tanchuk, V . Y . Application of associative neural networks for prediction of lipophilicity in alogps 2.1 program. Computer Methods and Programs in Biomedicine 77, 209–218 (2006).

33. Subramanian, G. & Kitchen, D. B. Computational models to predict blood–brain barrier permeation and cns activity. Journal of
Computer-Aided Molecular Design 17, 643–664 (2003).

32. Singh, M., Divakaran, R., Konda, L. S. K. & Kristam, R. A classification model for blood brain barrier penetration.

31. Li, H.

30. Gao, Z., Chen, Y., Cai, X. & Xu, R. Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and
drug indications. Bioinformatics 33, 901–908 (2017).

29. Andres, C. & Hutter, M. C. Cns permeability of drugs predicted by a decision tree. QSAR & Combinatorial Science 25, 305–309
(2006).

28. Roy, D., Hinge, V . K. & Kovalenko, A. 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 (2019).

27. Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A. & Tropsha, A. Qsar modeling of the blood–brain barrier permeability for diverse
organic compounds. Pharmaceutical Research 25, 1902–1914 (2008).

26. Chen, H., Vinwarter, S., Fridén, M., Antonsson, M. & Engkvist, O. In silico prediction of unbound brain-to-plasma concentration
ratio using machine learning algorithms. Journal of Molecular Graphics and Modelling 29, 985–995 (2011).

25. Konovalov, D. A., Coomans, D. & Deconinck, E. & Vander Heyden, Y. Benchmarking of qsr models for blood-brain barrier
permeation. Journal of Chemical Information and Modeling 47, 1648–1656 (2007).

24. Shayanfar, A., Soltani, S. & Jouyban, A. Prediction of blood–brain distribution: effect of ionization. Biological and Pharmaceutical
Bulletins 34, 266–271 (2011).

23. Vilaa, S., Chakrabortii, M. & Costanzii, S. Prediction of passive blood–brain partitioning: straightforward and effective classification
models based on in silico derived physicochemical descriptors. Journal of Molecular Graphics and Modelling 28, 899–903 (2010).

22. Toropov, A. A., Toropova, A. P ., Beeg, M., Gobbi, M. & Salmona, M. Qsr model for blood-brain barrier permeation. Journal of
Pharmacological and Toxicological Methods 88, 7–18 (2017).

21. Dichiara, M., Amata, B., Turnaturi, R., Marrazzo, A. & Amata, E. Tuning properties for blood–brain barrier permeation: A statistics-
based analysis. ACS Chemical Neuroscience 11, 34–44 (2019).

20. Buja, R., Struck-Lewicka, W., Kaliszni, M., Kaliszni, R. & Markuszewski, M. J. Blood–brain barrier permeability mechanisms in
view of quantitative structure–activity relationships (qsar). Journal of Pharmaceutical and Biomedical Analysis 108, 29–37 (2015).

19. Hemmateenejad, B., MRI, R., Safapour, M. A. & Mehdipour, A. R. Accurate prediction of the blood–brain partitioning of a large set
of solutes using ab initio calculations and genetic neural network modeling. Journal of Computational Chemistry 27, 1125–1135
(2006).

18. Valencia, C. Y . M. Chemical composition of DOC, 25B-NBOMe, 25C-NBOMe and In silico modeling of permeability to the blood-brain
barrier (BBB). Master’s thesis, Universidad Nacional de Colombia, Colombia (2017).
69. Radchenko, E. V., Dyabina, A. S. & Palyulin, V. A. Towards deep neural network models for the prediction of the blood–brain barrier permeability for diverse organic compounds. Molecules 25, 5901 (2020).
70. Hou, T. & Xu, X. Adme evaluation in drug discovery. 3. modeling blood–brain barrier partitioning using simple molecular descriptors. Journal of Chemical Information and Computer Sciences 43, 2137–2152 (2003).
71. Norinder, U. & Haeberlein, M. Computational approaches to the prediction of the blood–brain distribution. Advanced Drug Delivery Reviews 54, 291–313 (2002).
72. Sobanśka, A. W., Hekner, A. & Brzezińska, E. Rp-18 hplc analysis of drugs’ ability to cross the blood-brain barrier. Journal of Chemistry 2019 (2019).

Acknowledgements
PWA acknowledged funding from Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canada Research Chairs, Compute Canada, the Shared Hierarchical Academic Research Computing Network, and CANARIE for financial and computational support. The authors thank OpenEye Scientific Software, Inc. for providing academic license of OEChem Toolkit.

Author contributions
P.W.A. acquired the funding; P.W.A. and F.W.M. and conceived the study and wrote the manuscript. F.W.M., J.F.H. and X.Y. conducted the study and analyzed the results. All authors reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to P.W.A.

Reprints and permissions information is available at www.nature.com/reprints.

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

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 license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license 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 license, 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 metadata files associated with this article.

© The Author(s) 2021