Supporting Information

Computational Model to Predict the Fraction of Unbound Drug in the Brain

Tsuyoshi Esaki,#, Rikiya Ohashi#,‡, Reiko Watanabe#, Yayoi Natsume-Kitatani#,†, Hitoshi Kawashima#, Chioko Nagao#,†, Kenji Mizuguchi*,#,†*

[#] Laboratory of Bioinformatics, National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, 567-0085, Japan

[‡] Discovery Technology Laboratories, Mitsubishi Tanabe Pharma Corporation, 2-2-50 Kawagishi, Toda, Saitama, 335-8505, Japan

[†] Laboratory of In-silico Drug Design, Center of Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition, 7-6-8 Saito-Asagi, Ibaraki, Osaka, 567-0085, Japan

Corresponding Author

* (T.E.) E-mail: tsuyoshi-esaki@nibiohn.go.jp

* (K. M.) E-mail: kenji@nibiohn.go.jp

Table S1. List of the Compounds used in this Study.

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL1004	DOXYLAMINE	Training	0.038
CHEMBL1106	EPINASTINE	Training	0.603
CHEMBL111	RIMONABANT	Training	0.0004875
CHEMBL112	ACETAMINOPHEN	Training	0.832
CHEMBL1172	DESLORATADINE	Training	0.007
CHEMBL1185	ZOLMITRIPTAN	Training	0.537
CHEMBL1201201	METHAMPHETAMINE	Training	0.05
CHEMBL122	ROFECOXIB	Training	0.081
CHEMBL1237044	TRAMADOL	Training	0.356
CHEMBL125	MILTEFOSINE	Training	0.083
CHEMBL1256	ISOFLURANE	Training	0.087
CHEMBL1276947		Training	0.023
CHEMBL1276948		Training	0.032
CHEMBL1277126		Training	0.027
CHEMBL1277312		Training	0.04
CHEMBL1277678		Training	0.02
CHEMBL1277770		Training	0.031
CHEMBL1277935		Training	0.03
CHEMBL1278	NARATRIPTAN	Training	0.229
CHEMBL1286	LEVETIRACETAM	Training	1
CHEMBL129	ZIDOVUDINE	Training	0.95
CHEMBL131	PREDNISOLONE	Training	0.204
CHEMBL1324	TOLCAPONE	Training	0.00875
CHEMBL13280	FLUNITRAZEPAM	Training	0.219

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL134	CLONIDINE	Training	0.42
CHEMBL1380	ABACAVIR	Training	0.832
CHEMBL1428	NIMODIPINE	Training	0.013
CHEMBL1431	METFORMIN	Training	0.95
CHEMBL1434	MINOCYCLINE	Training	0.457
CHEMBL14370	REBOXETINE	Training	0.05
CHEMBL1457	HYDROCODONE	Training	0.465
CHEMBL146227		Training	0.24
CHEMBL1464	WARFARIN	Training	0.25375
CHEMBL1471	APREPITANT	Training	0.001375
CHEMBL1510	ELETRIPTAN	Training	0.055
CHEMBL159	VINBLASTINE	Training	0.005
CHEMBL16	PHENYTOIN	Training	0.120111111
CHEMBL160	CYCLOSPORINE	Training	0.025
CHEMBL1621	PALIPERIDONE	Training	0.111481818
CHEMBL1751	DIGOXIN	Training	0.214
CHEMBL177756	FLUORESCEIN	Training	0.42
CHEMBL1830693		Training	0.014
CHEMBL1830698		Training	0.0043
CHEMBL1830707		Training	0.0043
CHEMBL1830711		Training	0.016
CHEMBL2146883	COBIMETINIB	Training	0.0012
CHEMBL2151817		Training	0.01
CHEMBL21578	QUINIDINE	Training	0.0615
CHEMBL2164048		Training	0.003

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL2204445		Training	0.01
CHEMBL220492	TOPIRAMATE	Training	0.5285
CHEMBL2349466		Training	0.1
CHEMBL2349467		Training	0.0285
CHEMBL2349468		Training	0.006
CHEMBL2349469		Training	0.018
CHEMBL2349470		Training	0.097
CHEMBL2349471		Training	0.1
CHEMBL2349472		Training	0.003
CHEMBL2349475		Training	0.013
CHEMBL2349476		Training	0.034
CHEMBL2349478		Training	0.012
CHEMBL2349480		Training	0.22
CHEMBL2349490		Training	0.004
CHEMBL2349504		Training	0.03
CHEMBL24	ATENOLOL	Training	0.9
CHEMBL2403769		Training	0.031
CHEMBL2418359		Training	0.051
CHEMBL2418364		Training	0.023
CHEMBL260374	THIOPERAMIDE	Training	0.363
CHEMBL267894	AMOBARBITAL	Training	0.49
CHEMBL273575	NOMIFENSINE	Training	0.002
CHEMBL278020	NEOSTIGMINE	Training	0.95
CHEMBL281786	DIPRENORPHINE	Training	0.309
CHEMBL28218	BROMPERIDOL	Training	0.009

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL3127106		Training	0.09
CHEMBL3298273		Training	0.02
CHEMBL3331521		Training	0.14
CHEMBL334491	BUDIPINE	Training	0.018
CHEMBL3353290		Training	0.006
CHEMBL3357656		Training	0.037
CHEMBL3357661		Training	0.04
CHEMBL3359272		Training	0.015
CHEMBL3422018		Training	0.031
CHEMBL34259	METHOTREXATE	Training	0.89
CHEMBL3608680		Training	0.525
CHEMBL3608684		Training	0.436
CHEMBL3608688		Training	0.109
CHEMBL3608740		Training	0.604
CHEMBL3608741		Training	0.995
CHEMBL3617649		Training	0.032
CHEMBL363295	TERODILINE	Training	0.05375
CHEMBL3633720		Training	0.09
CHEMBL3633881		Training	0.16
CHEMBL3633943		Training	0.51
CHEMBL3634122		Training	0.029
CHEMBL3634340		Training	0.019
CHEMBL3746457		Training	0.055
CHEMBL3747116		Training	0.065
CHEMBL3799685		Training	0.001

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL3808485		Training	0.03
CHEMBL384467	DEXAMETHASONE	Training	0.38
CHEMBL424	SALICYLIC	Training	0.63
CHEMBL428647	PACLITAXEL	Training	0.0075
CHEMBL441	THIOPENTAL	Training	0.16
CHEMBL44657	ETOPOSIDE	Training	0.603
CHEMBL485	CODEINE	Training	0.62
CHEMBL492591		Training	1
CHEMBL493	BROMOCRIPTINE	Training	0.001875
CHEMBL502	DONEPEZIL	Training	0.10725
CHEMBL53463	DOXORUBICIN	Training	0.001
CHEMBL546	OXPRENOLOL	Training	0.28
CHEMBL56564	TROPISETRON	Training	0.05425
CHEMBL593	DELAVIRDINE	Training	0.022
CHEMBL596	FENTANYL	Training	0.071
CHEMBL596809		Training	0.027
CHEMBL597190		Training	0.006
CHEMBL607	MEPERIDINE	Training	0.054
CHEMBL608151		Training	0.014
CHEMBL610795		Training	0.001
CHEMBL629	AMITRIPTYLINE	Training	0.01
CHEMBL634	ALFENTANIL	Training	0.324
CHEMBL636	RIVASTIGMINE	Training	0.3755
CHEMBL639	AZELASTINE	Training	0.012375
CHEMBL649	NADOLOL	Training	0.78

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL651	METHADONE	Training	0.029
CHEMBL656	OXYCODONE	Training	0.66
CHEMBL657	DIPHENHYDRAMINE	Training	0.0472
CHEMBL658	SUFENTANIL	Training	0.034
CHEMBL70	MORPHINE	Training	0.7105
CHEMBL701	BACLOFEN	Training	0.98
CHEMBL71	CHLORPROMAZINE	Training	0.00293
CHEMBL72	DESIPRAMINE	Training	0.0115
CHEMBL7413	METHOHEXITAL	Training	0.331
CHEMBL74926	PHENSERINE	Training	0.065
CHEMBL7728	HEXOBARBITAL	Training	0.417
CHEMBL809	SERTRALINE	Training	0.001
CHEMBL81923	DEHYDROEVODIAMINE	Training	0.355
CHEMBL855	TRIPROLIDINE	Training	0.091
CHEMBL856	PRIMIDONE	Training	0.851
CHEMBL894	BUPROPION	Training	0.1689
CHEMBL9	NORFLOXACIN	Training	0.58
CHEMBL905	RIZATRIPTAN	Training	0.347
CHEMBL939	GEFITINIB	Training	0.191
CHEMBL963	OXYMORPHONE	Training	0.75
CHEMBL1000	CETIRIZINE	Test	0.309
CHEMBL1112	ARIPIPRAZOLE	Test	0.001125
CHEMBL1200733	DESFLURANE	Test	0.219
CHEMBL1595	DIHYDROCODEINE	Test	0.141
CHEMBL170	QUININE	Test	0.082

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL190	THEOPHYLLINE	Test	0.562
CHEMBL2057392		Test	0.009
CHEMBL2426616		Test	0.18
CHEMBL2431173		Test	0.016
CHEMBL278255	CLINAFLOXACIN	Test	0.264
CHEMBL28509	EDELFOSINE	Test	0.014
CHEMBL33	LEVOFLOXACIN	Test	0.84
CHEMBL3357669		Test	0.0135
CHEMBL3422010		Test	0.028
CHEMBL3527069		Test	0.0228
CHEMBL3527070		Test	0.004115
CHEMBL3617655		Test	0.0435
CHEMBL3617658		Test	0.033
CHEMBL3634123		Test	0.101
CHEMBL3634125		Test	0.044
CHEMBL3634341		Test	0.01
CHEMBL3634342		Test	0.039
CHEMBL3823424		Test	0.098
CHEMBL389621	HYDROCORTISONE	Test	0.178
CHEMBL407	FLUMAZENIL	Test	0.832
CHEMBL419296	ZOLANTIDINE	Test	0.023
CHEMBL500	PINDOLOL	Test	0.4
CHEMBL521982		Test	0.001
CHEMBL554	LAPATINIB	Test	0.011
CHEMBL70209	ZALTIDINE	Test	0.631

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL744	RILUZOLE	Test	0.011625
CHEMBL750	ZONISAMIDE	Test	0.43
CHEMBL814	FLUVOXAMINE	Test	0.0205
CHEMBL84	TOPOTECAN	Test	0.912
CHEMBL8809	RACLOPRIDE	Test	0.129
CHEMBL953	ENTACAPONE	Test	0.02275
CHEMBL1088	MESORIDAZINE	Wan	0.025
CHEMBL1113	AMOXAPINE	Wan	0.01
CHEMBL1233	CARISOPRODOL	Wan	0.335
CHEMBL128	SUMATRIPTAN	Wan	0.5435
CHEMBL1628227	DOXEPIN	Wan	0.025
CHEMBL21731	MAPROTILINE	Wan	0.006
CHEMBL41	FLUOXETINE	Wan	0.0028
CHEMBL42	CLOZAPINE	Wan	0.01065
CHEMBL445	NORTRIPTYLINE	Wan	0.00622
CHEMBL49	BUSPIRONE	Wan	0.2111
CHEMBL531	PERGOLIDE	Wan	0.027
CHEMBL567	PERPHENAZINE	Wan	0.004
CHEMBL621	TRAZODONE	Wan	0.0682
CHEMBL637	VENLAFAXINE	Wan	0.205
CHEMBL654	MIRTAZAPINE	Wan	0.065
CHEMBL660	AMANTADINE	Wan	0.1985
CHEMBL669	CYCLOBENZAPRINE	Wan	0.0046
CHEMBL715	OLANZAPINE	Wan	0.034
CHEMBL716	QUETIAPINE	Wan	0.025

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL741	LAMOTRIGINE	Wan	0.24675
CHEMBL796	METHYLPHENIDATE	Wan	0.195
CHEMBL831	LOXAPINE	Wan	0.011
CHEMBL85	RISPERIDONE	Wan	0.07076
CHEMBL896	HYDROXYZINE	Wan	0.011366667
CHEMBL911	ZOLPIDEM	Wan	0.15
CHEMBL95	TACRINE	Wan	0.142
CHEMBL972	SELEGILINE	Wan	0.064333333
CHEMBL1017	TELMISARTAN	Mateus	0.012875
CHEMBL103	PROGESTERONE	Mateus	0.046
CHEMBL1098	BUPIVACAINE	Mateus	0.2055
CHEMBL11	IMIPRAMINE	Mateus	0.035
CHEMBL114	SAQUINAVIR	Mateus	0.0029
CHEMBL115	INDINAVIR	Mateus	0.072
CHEMBL116	AMPRENAVIR	Mateus	0.091
CHEMBL13	METOPROLOL	Mateus	0.46
CHEMBL139	DICLOFENAC	Mateus	0.041
CHEMBL1477	CERIVASTATIN	Mateus	0.048
CHEMBL1484	NICARDIPINE	Mateus	0.005625
CHEMBL152067	TALINOLOL	Mateus	0.1495
CHEMBL163	RITONAVIR	Mateus	0.018222222
CHEMBL1790041	RANITIDINE	Mateus	0.955
CHEMBL266195	ALPRENOLOL	Mateus	0.057
CHEMBL267930	SPIPERONE	Mateus	0.037
CHEMBL295698	KETOCONAZOLE	Mateus	0.012

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL30	CIMETIDINE	Mateus	0.6425
CHEMBL374478	RIFAMPICIN	Mateus	0.13
CHEMBL415	CLOMIPRAMINE	Mateus	0.004
CHEMBL421	SULFASALAZINE	Mateus	0.063
CHEMBL46	ONDANSETRON	Mateus	0.049
CHEMBL498	CHLORPROPAMIDE	Mateus	0.68775
CHEMBL503	LOVASTATIN	Mateus	0.008
CHEMBL52440	DEXTROMETHORPHAN	Mateus	0.2
CHEMBL568	OXAZEPAM	Mateus	0.037
CHEMBL58	MITOXANTRONE	Mateus	0.0046
CHEMBL584	NELFINAVIR	Mateus	0.00336
CHEMBL6	INDOMETHACIN	Mateus	0.045333333
CHEMBL644	TRIMIPRAMINE	Mateus	0.007
CHEMBL661	ALPRAZOLAM	Mateus	0.16
CHEMBL6966	VERAPAMIL	Mateus	0.0425
CHEMBL79	LIDOCAINE	Mateus	0.2795
CHEMBL841	LOPERAMIDE	Mateus	0.0095
CHEMBL850	SPARFLOXACIN	Mateus	0.257
CHEMBL914	FEXOFENADINE	Mateus	0.078
CHEMBL998	LORATADINE	Mateus	0.002
CHEMBL108	CARBAMAZEPINE	Wan, Mateus	0.184
CHEMBL12	DIAZEPAM	Wan, Mateus	0.0405
CHEMBL27	PROPRANOLOL	Wan, Mateus	0.020666667
CHEMBL479	THIORIDAZINE	Wan, Mateus	0.000835
CHEMBL490	PAROXETINE	Wan, Mateus	0.003955

ChEMBL ID	Name [a]	Dataset [b]	Obs. value [c]
CHEMBL54	HALOPERIDOL	Wan, Mateus	0.0074
CHEMBL549	CITALOPRAM	Wan, Mateus	0.043818182
CHEMBL655	MIDAZOLAM	Wan, Mateus	0.02275
CHEMBL86	METOCLOPRAMIDE	Wan, Mateus	0.270583333

[a] Names were collected from ChEMBL using their ChEMBL ID. [b] Dataset used in this study (Training: 144 compounds, Test: 36 compounds, Wan: Wan's 36 compounds[1], Mateus: Mateus' 46 compounds[2]). [c] Obtained from previous studies.

Table S2. List of the 53 Descriptors Selected by Boruta.

Descriptor	Detail	meanImp [a]
SLogP[c]	Wildman-Crippen LogP	9.758
SMR_VSA7[c]	MOE MR VSA Descriptor 7 (3.05 <= x < 3.63)	6.482
AETA_alpha[c]	averaged ETA core count	4.875
khs.aaCH[b]	Counts the number of occurrences of the E-state fragments	4.665
AATSC0p[c]	averaged and centered moreau-broto autocorrelation of lag 0 weighted by polarizability	4.547
APC2D6_C_X[d]	Count of C-X at topological distance 6	4.249
APC2D7_C_N[d]	Count of C-N at topological distance 7	4.086
AATS5p[c]	averaged moreau-broto autocorrelation of lag 5 weighted by polarizability	4.078
ATSC0m[c]	centered moreau-broto autocorrelation of lag 0 weighted by mass	3.953
FPSA3[c]	fractional charged partial positive surface area (version 3)	3.934
tpsaEfficiency[b]	Polar surface area expressed as a ratio to molecular size	3.891
SubFPC274[d]	Count of aromatic substructure	3.872
ZMIC1[c]	1-ordered Z-modified information content	3.674
GATS1i[c]	geary coefficient of lag 1 weighted by ionization potential	3.599
AATS1p[c]	averaged moreau-broto autocorrelation of lag 1 weighted by polarizability	3.581
AATS2i[c]	averaged moreau-broto autocorrelation of lag 2 weighted by ionization potential	3.546
AATSC0v[c]	Count of C-Cl at topological distance 5	3.539
ETA_psi_1[c]	ETA psi	3.524

Descriptor	Detail	meanImp [a]
GATS1p[c]	geary coefficient of lag 1 weighted by polarizability	3.496
n6aRing[c]	6-membered aromatic ring count	3.488
APC2D9_C_X[d]	Count of C-X at topological distance 9	3.483
VR2_A[c]	Calculates atom additive logP and molar refractivity values as described by Ghose and Crippen	3.478
AATSC0m[c]	Polar surface area expressed as a ratio to molecular size	3.457
ZMIC2[c]	2-ordered Z-modified information content	3.397
ETA_eta_FL[c]	spectral moment from Distance matrix	3.355
AATSC0c[c]	averaged and centered moreau-broto autocorrelation of lag 0 weighted by gasteiger charge	3.308
AATS5v[c]	Calculates atom additive logP and molar refractivity values	3.307
PEOE_VSA6[c]	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)	3.279
AMID_O[c]	averaged molecular ID on O atoms	3.254
SubFPC307[d]	Count of chiral center specified substructure	3.190
MATS1i[c]	moran coefficient of lag 1 weighted by ionization potential	3.164
MDEC-22[b]	molecular distance edge between secondary C and secondary C	3.162
AATS0p[c]	averaged moreau-broto autocorrelation of lag 0 weighted by polarizability	3.148
THSA[b]	sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2	3.069
SaasC[c]	Count of dialkylether	3.050
MATS1se[c]	Count of C-Cl at topological distance 9	2.998

Descriptor	Detail	meanImp [a]
AATS0m[c]	Predicted logP based on the atom-type method	2.797
AATSC1c[c]	averaged and centered moreau-broto autocorrelation of lag 1 weighted by gasteiger charge	2.773
ETA_dEpsilon_B[c]	ETA delta epsilon (type: B)	2.687
AATSC1i[c]	averaged and centered moreau-broto autocorrelation of lag 1 weighted by ionization potential	2.674
AMR[b]	the Ghose-Crippen molar refractivity	2.571
APC2D8_C_X[d]	Count of C-X at topological distance 8	2.552
APC2D8_C_N[d]	Count of C-N at topological distance 8	2.522
RPCS[c]	relative positive charge surface area	2.491
RPSA[c]	relative polar surface area	2.448
APC2D7_C_X[d]	Count of C-X at topological distance 7	2.385
MDEC-23[b]	molecular distance edge between secondary C and tertiary C	2.349
VP-7[b]	Evaluates the Kier & Hall Chi path indices of order 7	2.314
APC2D5_C_Cl[d]	Count of C-Cl at topological distance 5	2.267
BCUTdv-1l[c]	first lowest eigenvalue of Burden matrix weighted by valence electrons	2.237
ATSC3dv[c]	Returns the number of atoms in the largest pi chain	2.204
ZMIC5[c]	5-ordered Z-modified information content	2.196
Mi[c]	A variety of descriptors combining surface area and partial charge informatio	2.111

Features used for principal component analysis (PCA) and the training models. [a]Mean importance of descriptors (meanImp) was calculated using Boruta. Descriptors generated using [b]CDK, [c]Mordred, and [d]PaDEL-Descriptor.

Table S3. Training set statistics of the final models.

Algorithm	R^2	RMSE
RF	0.955	0.205
GB	0.970	0.149
R-SVM	0.892	0.281
L-SVM	0.711	0.446
PLS	0.536	0.563

These scores were calculated using selected parameters of each machine learning algorithm for training set.

From these tables (Table S3 and Table 4 in the main text), the possibility of over fitting of our GB model was shown. From these tables (Table S3 and Table 4 in the main text), the possibility of over fitting of our GB model was shown, thus, more amount of data of $f_{\mathrm{u,brain}}$ are required to be published for model construction.

Table S4. Predictive Result of the Random Forest (RF) Method for All Descriptors.

Evaluation	R^2	*RMSE*
Out-of-bag [a]	0.556	0.551
Test set [b]	0.532	0.551

[a] Out-of-bag used instead of cross-validation; [b] Optimized model with mtry = 442.

The RF algorithm has a function to select important descriptors for training within the algorithm.[3] An $f_{\mathrm{u,brain}}$ predictive model was constructed with RF using the same training and test compounds used in previous models to confirm the predictive ability of this method for all or selected descriptors.

A total of 7,513 descriptors were generated with near-zero variance or high correlation descriptors filtered; 868 descriptors were retained for further analysis. As a result of the training conducted on these descriptors, 442 was selected as a suitable parameter for mtry, and statistical scores for evaluation were calculated using the test compounds (Table S1).

As a result of the comparison between all and selected descriptors, the RF model-trained descriptors selected by Boruta had slightly higher performance in both out-of-bag and test set validation. The descriptors chosen as the important features in RF were similar to the descriptors selected by Boruta (Table 2). Boruta is a wrapper algorithm with a combination of RF in the decision of importance for each descriptor. Using descriptors selected by Boruta for RF training did not affect our data.

Table S5. List of the 20 Most Important Descriptors Selected by Random Forest (RF).

Descriptor	Detail	Importance [a]
SLogP [c]	Wildman-Crippen LogP	100.00
SMR_VSA7 [c]	MOE MR VSA Descriptor 7 (3.05 <= x < 3.63)	42.46
APC2D6_C_X [d]	Count of C-X at topological distance 6	30.18
MDEC-22 [b]	molecular distance edge between secondary C and secondary C	25.89
FPSA3 [c]	fractional charged partial positive surface area (version 3)	25.73
GATS1p [c]	geary coefficient of lag 1 weighted by polarizability	25.09
AATSC0m [c]	Polar surface area expressed as a ratio to molecular size	24.75
AETA_alpha [c]	averaged ETA core count	23.62
khs.aaCH [b]	Counts the number of occurrences of the E-state fragments	22.99
AATSC0p [c]	averaged and centered moreau-broto autocorrelation of lag 0 weighted by polarizability	21.94
tpsaEfficiency [b]	Polar surface area expressed as a ratio to molecular size	21.74
ATSC0m [c]	centered moreau-broto autocorrelation of lag 0 weighted by mass	21.45
APC2D7_C_N [d]	Count of C-N at topological distance 7	21.24
JGI9	9-ordered mean topological charge	20.26
AATS5p [c]	averaged moreau-broto autocorrelation of lag 5 weighted by polarizability	20.25
APC2D9_C_X [d]	Count of C-X at topological distance 9	19.74
AATS1p [c]	averaged moreau-broto autocorrelation of lag 1 weighted by polarizability	19.62
AATSC1i [c]	averaged and centered moreau-broto autocorrelation of lag 1 weighted by ionization potential	19.58
SubFPC274 [d]	Count of aromatic substructure	19.58
ATSC6v [c]	centered moreau-broto autocorrelation of lag 6 weighted by	19.53

Descriptor	Detail	Importance [a]
	vdw volume	

The 868 filtered descriptors were used for descriptor selection and model building in RF training. [a]The importance of the descriptors was scaled, assuming the value of the most important descriptor as 100.00 in RF. All descriptors were generated using [b]CDK, [c]Mordred, and [d]PaDEL-Descriptor.

To select important descriptors, we employed Boruta and Random Forest in this study. In these methods, important descriptors are randomly selected and evaluated in each branch of trees. Table S4 contains the MDEC-22 (molecular distance edge between secondary C and secondary C) within the first 20 descriptors (in the top 4) in the importance score of 25.89. However, it was placed in the top 32 descriptors in the meanImp of 3.162 in Table S2.

The scores of importance and meanImp of MDEC-22 were approximately one of three of that of SLogP, which is the most important descriptor in both RF and Boruta selections. Thus, it seems MDEC-22 was randomly selected as having a higher importance than other descriptors in RF and accidentally had a lower order in Boruta selection. The values of MDEC-22 imply that they are reasonable for descriptor selection and were selected without bias.

References

1. Wan, H.; Åhman, M.; Holmen, A. G. Relationship Between Brain Tissue Partitioning and Microemulsion Retention Factors of CNS Drugs. *J. Med. Chem.*, **2009**, *52*, 1693-1700.

2. Mateus, A.; Matsson, P.; Artursson, P. A High-Throughput Cell-Based Method to Predict the Unbound Drug Fraction in the Brain. *J. Med. Chem.*, **2014**, *57*, 3005-3010.

3. Breiman, L. Random Forest, *Machine Learning*, **2001**, *45*, 5-32.