Supporting Information

pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures

Douglas E. V. Pires[1,2,‡,*], Tom L. Blundell[1], David B. Ascher[1,‡,*]

Table of Contents

Table S3. pkCSM prediction performance for anti-neoplastic drugs within evaluated data sets.

Figure S1. Regression analysis for Distribution predictors considering cross-validation schemes. Pearson's correlation coefficients and standard error are also shown.

Figure S2. Regression analysis for Toxicity predictors considering cross-validation schemes. Pearson's correlation coefficients and standard error are also shown.

Figure S3. Regression analysis for solubility prediction using the training and test data from the Solubility Challenge.

Figure S4. Regression analysis for Caco2 Permeability and Fraction Unbound prediction for macrocycles.

Predictive Models Implemented by pkCSM

Water Solubility

The water solubility of a compound (logS) reflects the solubility of the molecule in water at 25ºC. Lipid-soluble drugs are less well absorbed than water-soluble ones, especially when they are enteral. This model is built using experimental water solubility measurements of 1,708 molecules.

How to interpret the results:

The predicted water solubility of a compound is given as the logarithm of the molar concentration (log mol/L).

Caco-2 Permeability

The Caco-2 cell line is composed of human epithelial colorectal adenocarcinoma cells. The Caco-2 monolayer of cells is widely used as an *in vitro* model of the human intestinal mucosa to predict the absorption of orally administered drugs. This model is based on 674 drug like molecules with Caco-2 permeability values and predicts the logarithm of the apparent permeability coefficient (log Papp; log cm/s).

How to interpret the results:

A compound is considered to have a high Caco-2 permeability if it has a Papp > 8 x 10^{-6} cm/s. For the pkCSM predictive model, high Caco-2 permeability would translate in predicted values > 0.90.

Intestinal Absorption (Human)

The Intestine is normally the primary site for absorption of a drug from an orally administered solution. This method is built to predict the proportion of compounds that were absorbed through the human small intestine.

How to interpret the results:

For a given compound it predicts the percentage that will be absorbed through the human intestine. A molecule with an absorbance of less than 30% is considered to be poorly absorbed.

Skin Permeability

Skin permeability is a significant consideration for many consumer products efficacy, and of interest for the development of transdermal drug delivery. This predictor was built using 211 compounds whose in vitro human skin permeability has been measured

How to interpret the results:

It predicts whether if given compound is likely to be skin permeable, expressed as the skin permeability constant logKp (cm/h). A compound is considered to have a relatively low skin permeability if it has a logKp > -2.5.

P-glycoprotein substrate

The P-glycoprotein is an ATP-binding cassette (ABC) transporter. It functions as a biological barrier by extruding toxins and xenobiotics out of cells. P-glycoprotein transport screening is performed using transgenic mdr knockout mice and *in vitro* cell systems. This model was built using 332 compounds that have been characterised for their ability to be transported by Pgp.

How to interpret the results:

The model predicts whether a given compound is likely to be a substrate of Pgp or not.

P-glycoprotein I and II inhibitors

Modulation of P-glycoprotein mediated transport has significant pharmacokinetic implications for Pgp substrates, which may either be exploited for specific therapeutic advantages or result in

contraindications. This predictive models were build using 1,273 and 1,275 compounds that have been characterised for their ability to inhibit P-glycoprotein I and P-glycoprotein II transport, respectively.

How to interpret the results:

The predictor will determine is a given compound is likely to be a P-glycoprotein I/II inhibitor.

VDss (Human)

The steady state volume of distribution (VDss) is the theoretical volume that the total dose of a drug would need to be uniformly distributed to give the same concentration as in blood plasma. The higher the VD is, the more of a drug is distributed in tissue rather than plasma. It can be affected by renal failure and dehydration. This predictive model was built using the calculated steady state volume of distribution (VDss) in humans from 670 drugs. The predicted logarithm of VDss of a given compound is given as the log L/kg.

How to interpret the results:

VDss is considered low if below 0.71 L/kg (log VDss < -0.15) and high if above 2.81 L/kg (log VDss > 0.45).

Fraction Unbound (Human)

Most drugs in plasma will exist in equilibrium between either an unbound state or bound to serum proteins. Efficacy of a given drug may be affect by the degree to which it binds proteins within blood, as the more that is bound the less efficiently it can traverse cellular membranes or diffuse. This predictive model was built using the measured free proportion of 552 compounds in human blood (Fu).

How to interpret the results:

For a given compound the predicted fraction that would be unbound in plasma will be calculated.

Blood Brain Barrier permeability

The brain is protected from exogenous compounds by the blood-brain barrier (BBB). The ability of a drug to cross into the brain is an important parameter to consider to help reduce side effects and toxicities or to improve the efficacy of drugs whose pharmacological activity is within the brain. Blood-brain permeability is measured *in vivo* in animals models as logBB, the logarithmic ratio of brain to plasma drug concentrations. This predictive model was built using 320 compounds whose logBB has been experimentally measured.

How to interpret the results:

For a given compound, a logBB > 0.3 considered to readily cross the blood-brain barrier while molecules with logBB < −1 are poorly distributed to the brain.

CNS permeability

Measuring blood brain permeability can difficult with confounding factors. The blood-brain permeability-surface area product (logPS) is a more direct measurement. It is obtained from *in situ* brain perfusions with the compound directly injected into the carotid artery. This lacks the systemic distribution effects which may distort brain penetration. This predictive model was built using 153 compounds whose logPS has been experimentally measured.

How to interpret the results:

Compounds with a logPS > -2 are considered to penetrate the Central Nervous System (CNS), while those with logPS < -3 are considered as unable to penetrate the CNS.

CYP2D6/CYP3A4 substrate

The cytochrome P450's are responsible for metabolism of many drugs. However inhibitors of the P450's can dramatically alter the pharmacokinetics of these drugs. It is therefore important to assess whether a given compound is likely to be a cytochrome P450 substrate. The two main isoforms responsible for drug

metabolism are 2D6 and 3A4. These models were built using 671 compounds whose metabolism by each cytochrome P450 isoform has been measured.

How to interpret the results:

The predictor will assess whether a given molecule is likely to be metabolised by either P450.

Cytochrome P450 inhibitors

Cytochrome P450 is an important detoxification enzyme in the body, mainly found in the liver. It oxidises xenobiotics to facilitate their excretion. Many drugs are deactivated by the cytochrome P450's, and some can be activated by it. Inhibitors of this enzyme, such as grapefruit juice, can affect drug metabolism and are contraindicated. It is therefore important to assess a compound's ability to inhibit the cytochrome P450. Models for different isoforms were built (CYP1A2/CYP2C19/CYP2C9/CYP2D6/CYP3A4) using from over 14000 to 18000 compounds whose ability to inhibit the cytochrome P450 has been determined. A compound is considered to be a cytochrome P450 inhibitor if the concentration required to lead to 50% inhibition is less than 10 uM.

How to interpret the results:

The predictors will assess a given molecule to determine whether it is likely going to be a cytochrome P450 inhibitor, for a given isoform.

Total Clearance

Drug clearance is measured by the proportionality constant CLtot, and occurs primarily as a combination of hepatic clearance (metabolism in the liver and biliary clearance) and renal clearance (excretion via the kidneys). It is related to bioavailability, and is important for determining dosing rates to achieve steady-state concentrations. This predictor was built using the total clearance data for 398 compounds.

How to interpret the results:

The predicted total clearance log(CLtot) of a given compound is given in log(ml/min/kg).

Renal OCT2 substrate

Organic Cation Transporter 2 is a renal uptake transporter that plays an important role in disposition and renal clearance of drugs and endogenous compounds. OCT2 substrates also have the potential for adverse interactions with coadministered OCT2 inhibitors. Assessing a candidate's potential to be transported by OCT2 provides useful information regarding not only its clearance but potential contraindications. This model was built using 906 compounds whose transport by OCT2 has been experimentally measured.

How to interpret the results:

The predictor will assess whether a given molecule is likely to be an OCT2 substrate.

AMES toxicity

The Ames test is a widely employed method to assess a compounds mutagenic potential using bacteria. A positive test indicates that the compound is mutagenic and therefore may act as a carcinogen. This predictive model was built on the results of over 8,000 compounds Ames tests.

How to interpret the results:

It predicts whether a given compound is likely to be Ames positive and hence mutagenic.

Maximum Tolerated Dose (Human)

The maximum recommended tolerated dose (MRTD) provides an estimate of the toxic dose threshold of chemicals in humans. The model is trained using 1222 experimental data points from human clinical trials and predicts the logarithm of the MRTD (log mg/kg/day). This will help guide the maximum recommended starting dose for pharmaceuticals in phase I clinical trials, which are currently based on extrapolations from animal data.

How to interpret the results:

For a given compound, a MRTD of less than or equal to 0.477 log(mg/kg/day) is considered low, and high if greater than 0.477 log(mg/kg/day).

hERG I and II Inhibitors

Inhibition of the potassium channels encoded by hERG (human ether-a-go-go gene) are the principal causes for the development of acquire long QT syndrome - leading to fatal ventricular arrhythmia. Inhibition of hERG channels has resulted in the withdrawal of many substances from the pharmaceutical market. These predictors were built using hERG I and II inhibition information for 368 and 806 compounds, respectively.

How to interpret the results:

The predictor will determine if a given compound is likely to be a hERG I/II inhibitor.

Oral Rat Acute Toxicity (LD50)

It is important to consider the toxic potency of a potential compound. The lethal dosage values (LD50) are a standard measurement of acute toxicity used to assess the relative toxicity of different molecules. The LD50 is the amount of a compound given all at once that causes the death of 50% of a group of test animals.

How to interpret the results:

The model was built on over 10000 compounds tested in rats and predicts the LD50 (in mol/kg).

Oral Rat Chronic Toxicity

Exposure to low-moderate doses of chemicals over long periods of time is of significant concern in many treatment strategies. Chronic studies aim to identify the lowest dose of a compound that results in an

observed adverse effect (LOAEL), and the highest dose at which no adverse effects are observed (NOAEL). This predictor was built using the LOAEL results from 445 compounds.

How to interpret the results:

For a given compound, the predicted log Lowest Observed Adverse Effect (LOAEL) in log(mg/kg_bw/day) will be generated. The LOAEL results need to be interpreted relative to the bioactive concentration and treatment lengths required.

Hepatotoxicity

Drug-induced liver injury is a major safety concern for drug development and a significant cause of drug attrition. This predictor was built using the liver associated side effects of 531 compounds observed in humans. A compound was classed as hepatotoxic if it had at least one pathological or physiological liver event which is strongly associated with disrupted normal function of the liver.

How to interpret the results:

It predicts whether a given compound is likely to be associated with disrupted normal function of the liver.

Skin Sensitisation

Skin sensitisation is a potential adverse effect for dermally applied products. The evaluation of whether a compound, that may encountered the skin, can induce allergic contact dermatitis is an important safety concern. This predictor was built using 254 compounds which have been evaluated for their ability to induce skin sensitisation.

How to interpret the results:

It predicts whether a given compound is likely to be associated with skin sensitisation.

T. Pryiformis toxicity

T. Pyriformis is a protozoa bacteria, with its toxicity often used as a toxic endpoint. This method was build using the concentration of 1,571 compounds required to inhibit 50% of growth (IGC50).

How to interpret the results:

For a given compound, the pIGC50 (negative logarithm of the concentration required to inhibit 50% growth in log ug/L) is predicted, with a value > -0.5 log ug/L is considered toxic.

Minnow toxicity

The lethal concentration values (LC50) represent the concentration of a molecule necessary to cause the death of 50% of the Flathead Minnows. This predictive model was built on LC50 measurements for 554 compounds.

How to interpret the results:

For a given compound, a log LC50 will be predicted. LC50 values below 0.5 mM (log LC50 < -0.3) are regarded as high acute toxicity.

Experimental Methods

Data sets

The datasets used are composed by small-molecules represented as SMILES strings with their respective experimental pharmacokinetic or toxicity measurement. In total, 30 datasets of different sizes, ranging from a few hundreds to over 18,000 compounds were collected from the literature.

The main sources were the work of Cheng and colleagues[1], the PKKB database[2], the work carried out by Obach and colleagues[3] on pharmacokinetics in humans for 670 drugs, the works on Central Nervous System permeability by Suenderhauf and colleagues[4] and Yan and colleagues[5] and the database of the FDA Maximum Recommended Daily Dose (MRTD) use in [6].

Evaluation data sets containing macrocyclic compounds were also identified from this initial pool by substructure search via a SMARTS query (available as Supplementary Material) which searches for rings with twelve of more atoms. Sufficient macrocycles were found for Caco2 permeability (24 compounds), Fraction Unbound (22 compounds) and for Cytochrome P450 inhibition (over 200 compounds).

A list of anti-neoplastic drugs was obtained from Drug Bank[7] using the 'antineoplastic agents' mesh term. Their SMILES were matched (using RDKit) to existing compounds on the evaluated data sets.

A complete view of the datasets used in this work can be obtained in Table S2.

Training and evaluating models

The qualitative predictions (classification tasks) were done by two different algorithms, Random Forest[8] and Logistic Regression[9]. The quantitative predictions (regression tasks) were also done by two different algorithms, Gaussian Processes[10] and Model Tree Regression[11]. The best performing predictor in each

task was chosen. The Weka toolkit was used for training and testing the models.

The usefulness and reliability of pkCSM was evaluated using different external data sets and cross-validation protocols and by comparing to the current leading approaches available for each predictive type. The description of the evaluation set up can be found in Table S2 of Supplementary Material.

Website Design and Implementation

pkCSM provides a user-friendly and quick web interface to generate ADMET predictions for up to 100 given chemical compounds at a time, developed using cutting edge frameworks (the front-end uses Bootstrap 2.0 and the back-end was implemented in Python, using Flask (0.10.1). The 2D chemical structures depiction are generated by RDkit.

Toxicophore SMARTS Queries

O=N(~O)a

a[NH2]

a[N;X2]=O

CO[N;X2]=O

N[N;X2]=O

O1[c,C]-[c,C]1

C1NC1

N=[N+]=[N-]

C=[N+]=[N-]

N=N-N

c[N;X2]!@;=[N;X2]c

[OH,NH2][N,O]

[OH]Na

[Cl,Br,I]C

[Cl,Br,I]C=O

[N,S]!@[C;X4]!@[CH2][Cl,Br,I]

[cH]1[cH]ccc2c1c3c(cc2)cc[cH][cH]3

[cH]1cccc2c1[cH][cH]c3c2ccc[cH]3

[$([C,c]OS(=O)(=O)O!@[c,C]),$([c,C]S(=O)(=O)O!@[c,C])]

O=N(~O)N

[$(O=[CH]C=C),$(O=[CH]C=O)]

[N;v4]#N

O=C1CCO1

[CH]=[CH]O

[NH;!R][NH;R]a

[CH3][NH]a

aN([$([OH]),$(O*=O)])[$([#1]),$(C(=O)[CH3]),$([CH3]),$([OH]),$(O*=O)]

a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~a~2

a1~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3

a1~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3

a1~a~a~a~a2~a~1~a3~a(~a~2)~a~a~a~a~3

a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~3

a1~a~a~a~a2~a~1~a~a3~a(~a~2)~a~a~a~a~3

a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~3

a1~a~a~a~a2~a~1~a~a~a3~a~2~a~a~a~a~3

a13~a~a~a~a2~a1~a(~a~a~a~3)~a~a~2

Macrocycles SMARTS Query

[r;!r3;!r4;!r5;!r6;!r7;!r8;!r9;!r10;!r11]

Case Studies

Case Study 1: The Solubility Challenge

The solubility of a compound is a key physicochemical property, important in both chemistry and biology, and influences its pharmacokinetic behavior. It has been regarded as a difficult property to predict, with many computational models suggested to be over fitted, with large errors and hence low reliability[12]. To address this, Llinas and colleagues proposed a solubility challenge[13], to which more than 100 entries were submitted[14]. To compare the predictive performance of our approach, we subjected it to this test.

On the training set, pkCSM managed a Pearson correlation of 0.818 (0.911 after 10% of the outliers were removed, and a standard error of 0.846 (0.558 after 10% of the outliers were removed) (Figure S3). Using the trained model to analyze the test set, pkCSM achieved a correlation of 0.733 (0.879 after removal of 10% of outliers) and an standard error of 0.945 (0.653 after 10% of the outliers were removed) (Figure S3). By Fisher r−to−z transformation ($p<0.05$), pkCSM was more accurate than the entries presented by Hopfinger and collegues[14] or the final model presented by Hewitt et al[15] which obtained a correlation of 0.740 on the training set, and 0.510 on the test set, with a standard error of 0.95. This highlights the strength of the molecular signature approach presented here.

Case Study 2: Predicting pharmacokinetic properties of macrocycles

Macrocylces are defined as compounds with ring structures of 12 or more atoms. Their ability to target previously undruggable sites, including protein interfaces, has aroused significant interest, leading to already over 70 macrocyclic drugs in therapeutic use. However, macrocycles do not appear to conform to conventional metrics such as Lipinski's Rule of Five[16, 17]. Villar and colleagues have been able to extract rules for the development of oral macrocycles, confirming that these do differ significantly from other small molecules[18]. This may be due to the presence of internally-satisfied H-bonds, which were identified as a potential exception by Lipinski.

Within the databases used to train pkCSM, three models had experimental data for sufficient numbers of macrocycles. These included Caco2 permeability, Cytochrome P450 inhibition and Fraction Unbound. Considering their increasing importance and unique properties, we devised a study case to assess the performance of pkCSM on predicting properties of macrocycles identified in the datasets used to train pkCSM. Since the pharmacokinetic properties of macrocycles are known to not obey the same rules as the majority of drug like molecules, this was expected to challenge the limits of the pkCSM signatures. Despite the difference in ideal 'drug-like' properties between most drugs and the macrocylces, pkCSM was able to predict a broad range of the macrocycle pharmacokinetic properties.

There were approximately 20 macrocycles that had been experimentally characterized, and with a broad distribution, in the Caco2 and fraction unbound datasets. pkCSM was able to predict the Caco2 absorption of macrocycles (R^2=0.912, σ=0.305; right graph of Figure S4) and the fraction that would be unbound in plasma (R^2=0.922, σ=0.117; left graph of Figure S4), results compatible with the cross-validation performances obtained (Table 1).

There were also 200-300 macrocycles present in each of the Cytochrome P450 inhibition datasets. While pkCSM performed extremely well in its ability to classify the macrocycles according to their ability to inhibit the P450 subtypes (accuracy of 87%-98%, comparable with the cross-validation accuracy of 84%-88%), the macrocycles were poorly distributed across the two classes as the majority were not P450

inhibitors.

Case Study 3: Analyzing anti-neoplastic drugs

Anti-neoplastic drugs, by the nature of their mechanism of action, is one class of drugs associated with significant side effects and a narrow therapeutic window to balance their activity (dictated by their pharmacokinetics and pharmacodynamics) and toxicity. While this can be mitigated, for example by the use of drug carriers[19] or chemical modifications[20-22], it is still a serious concern during the drug development process and a significant cause of attrition during clinical trials.

Within the datasets used to train pkCSM there was well characterized data for a number of clinically used anti-neoplastic drugs. In a similar approach to our analysis of macrocylces above, we used this data to evaluate the performance of pkCSM on these drugs.

Table S3 summarizes the performance of pkCSM classification and regression models on predicting ADMET properties for the identified chemotherapeutics. Accuracies of over 80% for the majority of the classification models were achieved and correlation coefficients ranging from 0.67 (for Oral Rat Acute Toxicity) to 0.92 (for Fraction Unbound/Human), showcasing the efficacy of the proposed methodology.

Tables

Table S1: List of molecular properties calculated using the RDKit cheminformatics toolkit and used for training the predictive models.

Property	**Data type**
Molecular Weight	Real
Heavy Atom count	Integer
LogP	Real
Heteroatoms count	Integer
Rotatable Bonds count	Integer
Ring count	Integer
TPSA[a]	Real
Labute ASA[b]	Real
Fluorine atom Count	Integer
Toxicophore [1-36][c]	Binary vector
Pharmacophore count[d]	Integer vector

[a] Topological polar surface area[23].

[b] Labute ASA[24].

[c] Obtained via SMARTS queries proposed in[25].

[d] Six classes are considered: Hydrogen acceptor, hydrogen donor, aromatic ,hydrophobe, negative ionizable and positive ionizable.

Table S2: Description of datasets used on building the pkCSM platform and validation protocols employed.

Class	Dataset	Size	Task	Validation	Reference
Absorption	Water Solubility	1708	Reg.	5-fold cv.	[1]
Absorption	Caco2 permeability	674	Reg.	5-fold cv. & Train/test	[1]
Absorption	Human Intestinal Absorption (HIA)	552	Reg.	Train/test	[1]
Absorption	Skin Permeability	186	Reg.	5-fold cv.	[26]
Absorption	P-glycoprotein Substrate	332	Class.	5-fold cv.	[1]
Absorption	P-glycoprotein Inhibitor I	1273	Class.	5-fold cv.	[1]
Absorption	P-glycoprotein Inhibitor II	1275	Class.	5-fold cv.	[1]
Distribution	Human Volume of Distribution - (VDss)	670	Reg.	Leave-one-out cv. & Train/test	[3]
Distribution	Human Fraction Unbound (FU)	670	Reg.	Leave-one-out cv. & Train/test	[3]
Distribution	BBB permeability	320	Reg.	10-fold cv.	[4]
Distribution	CNS permeability	153	Reg.	10-fold cv.	[5]
Metabolism	CYP450 2D6 Substrate	671	Class.	5-fold cv.	[1]
Metabolism	CYP450 3A4 Substrate	671	Class.	5-fold cv.	[1]
Metabolism	CYP450 1A2 Inhibitor	14903	Class.	5-fold cv.	[1]
Metabolism	CYP450 C19 Inhibitor	14576	Class.	5-fold cv.	[1]
Metabolism	CYP450 2C9 Inhibitor	14709	Class.	5-fold cv.	[1]
Metabolism	CYP450 2D6 Inhibitor	14741	Class.	5-fold cv.	[1]
Metabolism	CYP450 3A4 Inhibitor	18561	Class.	5-fold cv.	[1]
Excretion	Total Clearance	503	Reg.	Train/test	[27]
Excretion	Renal Organic Cation Transporter	906	Class.	5-fold cv.	[1]
Toxicity	AMES Toxicity	8445	Class.	5-fold cv.	[1, 28]
Toxicity	Max. Recommended Therapeutic Dose (MRTD)	1220	Reg.	10-fold cv.	[6]
Toxicity	hERG I Inhibitor	368	Class.	5-fold cv.	[1]
Toxicity	hERG II Inhibitor	806	Class.	5-fold cv.	[1]
Toxicity	RAT LD50	10207	Reg.	5-fold cv.	[1]
Toxicity	Oral Rat Chronic Toxicity (LOAEL)	445	Reg.	Leave-one-out cv.	[29]
Toxicity	Hepatotoxicity	531	Class.	5-fold cv.	[30]
Toxicity	Skin Sensitisation	234	Class.	5-fold cv.	[31]
Toxicity	*T.Pyriformis* Toxicity (TPT)	1571	Reg.	5-fold cv.	[1]
Toxicity	Minnow toxicity (FHMT)	554	Reg.	5-fold cv.	[1]

Table S3. pkCSM prediction performance for anti-neoplastic drugs within evaluated data sets.

Dataset	#Drugs	Success rate*
AMES Toxicity	15	93.3%
CYP1A2 Inhibitor	27	85.2%
CYP2C19 Inhibitor	26	80.8%
CYP2C9 Inhibitor	29	82.8%
CYP2D6 Inhibitor	30	90.0%
CYP3A4 Inhibitor	37	89.2%
CYP2D6 Substrate	10	80.0%
CYP3A4 Substrate	10	70.0%
hERG-II Inhibitor	12	83.3%
Hepatotoxicity	12	100%
Fraction Unbound Human (Fu)	13	0.916
Steady State Volume of Distribution (VDss)	12	0.759
Total Clearance	10	0.703
Oral Rat Accute Toxicity (LD50)	7	0.668

*Accuracy for classification tasks and Pearson's correlation coefficient for regression tasks.

Figures

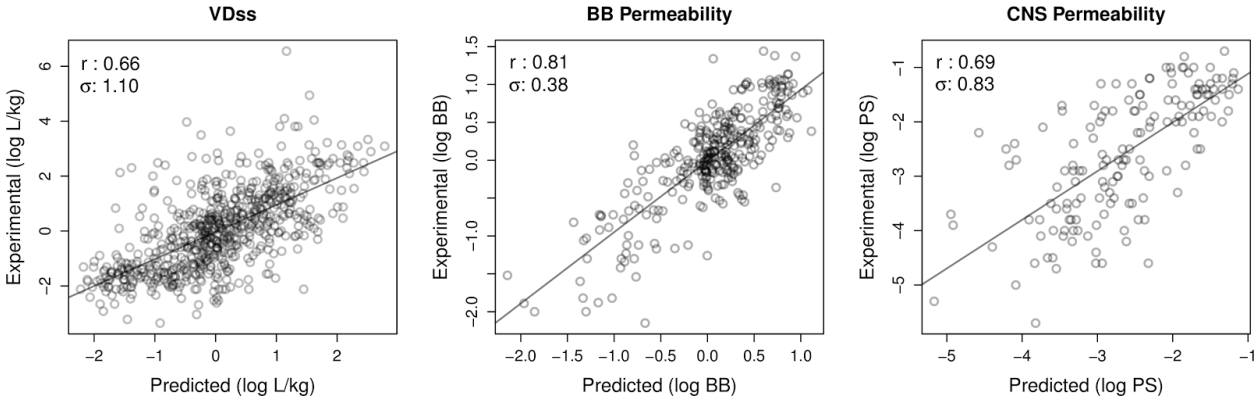

Figure S1. Regression analysis for Distribution predictors considering cross-validation schemes. Pearson's correlation coefficients and standard error are also shown.

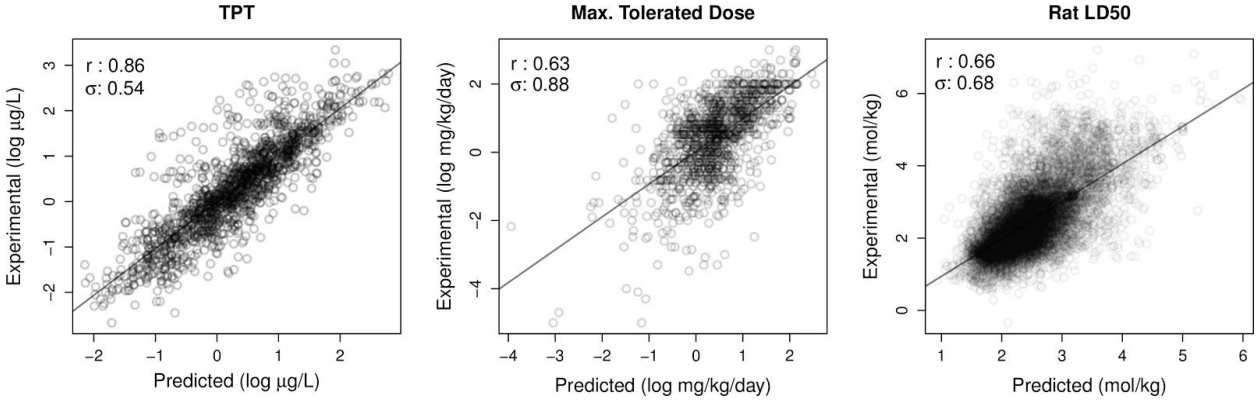

Figure S2. Regression analysis for Toxicity predictors considering cross-validation schemes. Pearson's correlation coefficients and standard error are also shown.

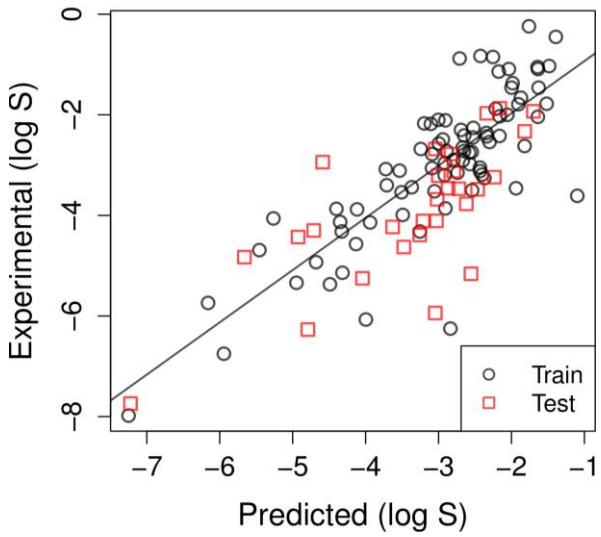

Figure S3. Regression analysis for solubility prediction using the training and test data from the Solubility Challenge[13].

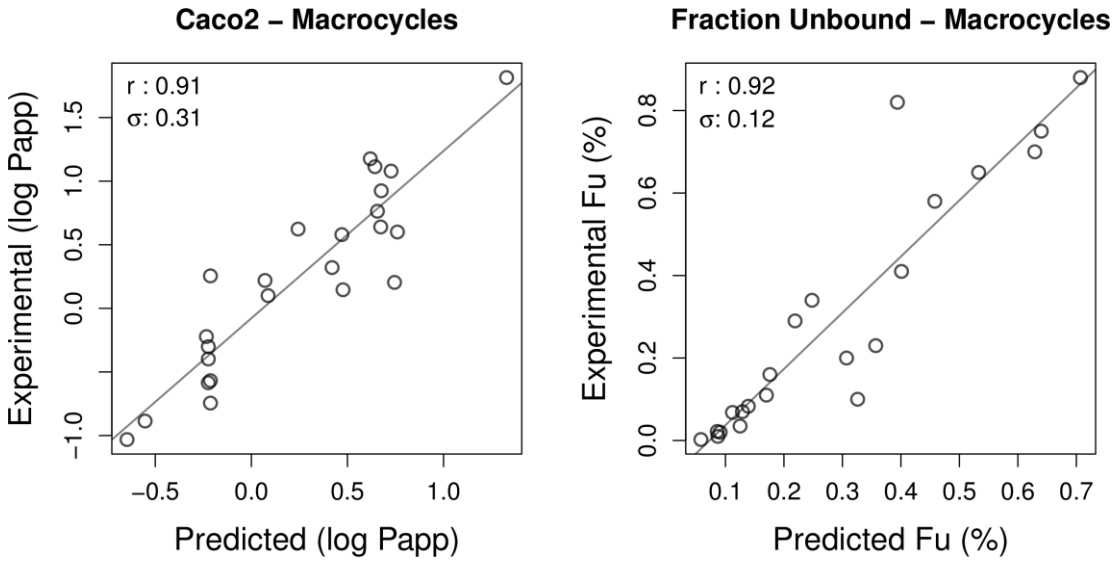

Figure S4. Regression analysis for Caco2 Permeability and Fraction Unbound prediction for macrocycles.

References

1. 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. *Journal of chemical information and modeling* **2012,** 52, 3099-105.

2. Cao, D.; Wang, J.; Zhou, R.; Li, Y.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 11. PharmacoKinetics Knowledge Base (PKKB): a comprehensive database of pharmacokinetic and toxic properties for drugs. *Journal of chemical information and modeling* **2012,** 52, 1132-7.

3. Obach, R. S.; Lombardo, F.; Waters, N. J. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. *Drug metabolism and disposition: the biological fate of chemicals* **2008,** 36, 1385-405.

4. Suenderhauf, C.; Hammann, F.; Huwyler, J. Computational prediction of blood-brain barrier permeability using decision tree induction. *Molecules* **2012,** 17, 10429-45.

5. Yan, A.; Liang, H.; Chong, Y.; Nie, X.; Yu, C. In-silico prediction of blood–brain barrier permeability. *SAR and QSAR in Environmental Research* **2012,** 24, 61-74.

6. Matthews, E. J.; Kruhlak, N. L.; Benz, R. D.; Contrera, J. F. Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data. *Current drug discovery technologies* **2004,** 1, 61-76.

7. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **2014,** 42, D1091-7.

8. Breiman, L. Random Forests. *Machine Learning* **2001,** 45.1, 5-32.

9. Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*. Wiley: 2004.

10. Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press: 2005.

11. Quinlan, J. R. Learning with continuous classes. *Proceedings of the 5th Australian joint Conference on Artificial Intelligence.* **1992,** 92.

12. Dearden, J. C. In silico prediction of aqueous solubility. *Expert opinion on drug discovery* **2006,** 1, 31-52.

13. Llinas, A.; Glen, R. C.; Goodman, J. M. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *Journal of chemical information and modeling* **2008,** 48, 1289-303.

14. Hopfinger, A. J.; Esposito, E. X.; Llinas, A.; Glen, R. C.; Goodman, J. M. Findings of the challenge to predict aqueous solubility. *Journal of chemical information and modeling* **2009,** 49, 1-5.

15. Hewitt, M.; Cronin, M. T.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In silico prediction of aqueous solubility: the solubility challenge. *Journal of chemical information and modeling* **2009,** 49, 2572-87.

16. Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chemistry & biology* **2014,** 21, 1115-42.

17. Giordanetto, F.; Kihlberg, J. Macrocyclic drugs and clinical candidates: what can medicinal chemists learn from their properties? *Journal of medicinal chemistry* **2014,** 57, 278-95.

18. Villar, E. A.; Beglov, D.; Chennamadhavuni, S.; Porco, J. A., Jr.; Kozakov, D.; Vajda, S.; Whitty, A. How proteins bind macrocycles. *Nature chemical biology* **2014,** 10, 723-31.

19. Kaminskas, L. M.; McLeod, V. M.; Ascher, D. B.; Ryan, G. M.; Jones, S.; Haynes, J. M.; Trevaskis, N. L.; Chan, L. J.; Sloan, E. K.; Finnin, B. A.; Williamson, M.; Velkov, T.; Williams, E. D.; Kelly, B. D.; Owen, D. J.; Porter, C. J. Methotrexate-conjugated PEGylated dendrimers show differential patterns of deposition and activity in tumor-burdened lymph nodes after intravenous and subcutaneous administration in rats. *Molecular pharmaceutics* **2015,** 12, 432-43.

20. Chan, L. J.; Bulitta, J. B.; Ascher, D. B.; Haynes, J. M.; McLeod, V. M.; Porter, C. J.; Williams, C. C.; Kaminskas, L. M. PEGylation Does Not Significantly Change the Initial Intravenous or Subcutaneous Pharmacokinetics or Lymphatic Exposure of Trastuzumab in Rats but Increases Plasma Clearance after Subcutaneous Administration. *Molecular pharmaceutics* **2015,** 12, 794-809.

21. Kaminskas, L. M.; Ascher, D. B.; McLeod, V. M.; Herold, M. J.; Le, C. P.; Sloan, E. K.; Porter, C. J. PEGylation of interferon alpha2 improves lymphatic exposure after subcutaneous and intravenous

administration and improves antitumour efficacy against lymphatic breast cancer metastases. *Journal of controlled release : official journal of the Controlled Release Society* **2013,** 168, 200-8.

22. Landersdorfer, C. B.; Caliph, S. M.; Shackleford, D. M.; Ascher, D. B.; Kaminskas, L. M. PEGylated Interferon Displays Differences in Plasma Clearance and Bioavailability Between Male and Female Mice and Between Female Immunocompetent C57Bl/6J and Athymic Nude Mice. *Journal of pharmaceutical sciences* **2015**.

23. Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of medicinal chemistry* **2000,** 43, 3714-7.

24. Labute, P. A widely applicable set of descriptors. *Journal of molecular graphics & modelling* **2000,** 18, 464-77.

25. Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry* **2005,** 48, 312-20.

26. Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. *Toxicol Appl Pharmacol* **2015**.

27. Yap, C. W.; Li, Z. R.; Chen, Y. Z. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. *Journal of molecular graphics & modelling* **2006,** 24, 383-95.

28. Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ Res* **2008,** 19, 495-524.

29. Mazzatorta, P.; Estevez, M. D.; Coulet, M.; Schilter, B. Modeling oral rat chronic toxicity. *Journal of chemical information and modeling* **2008,** 48, 1949-54.

30. Fourches, D.; Barnes, J. C.; Day, N. C.; Bradley, P.; Reed, J. Z.; Tropsha, A. Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chem Res Toxicol* **2010,** 23, 171-83.

31. Alves, V. M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C. H.; Tropsha, A. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol Appl Pharmacol* **2015**.