JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

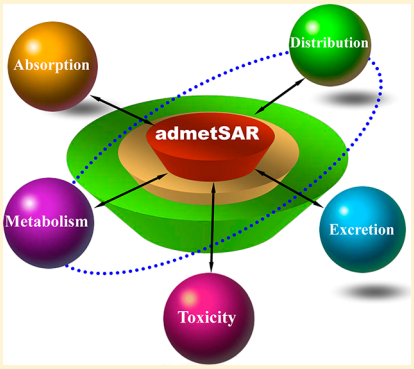
admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties

Feixiong Cheng, Weihua Li, Yadi Zhou, Jie Shen, Zengrui Wu, Guixia Liu, Philip W. Lee, and Yun Tang*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

Supporting Information

ABSTRACT: Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties play key roles in the discovery/development of drugs, pesticides, food additives, consumer products, and industrial chemicals. This information is especially useful when to conduct environmental and human hazard assessment. The most critical rate limiting step in the chemical safety assessment workflow is the availability of high quality data. This paper describes an ADMET structure−activity relationship database, abbreviated as admetSAR. It is an open source, text and structure searchable, and continually updated database that collects, curates, and manages available ADMET-associated properties data from the published literature. In admetSAR, over 210 000 ADMET annotated data points for more than 96 000 unique compounds with 45 kinds of ADMET-associated properties, proteins, species, or organisms have been carefully curated from a large number of diverse literatures. The database provides a user-friendly interface to query a specific chemical profile, using either CAS registry number, common name, or structure similarity. In addition, the database includes 22 qualitative classification and 5 quantitative regression models with highly predictive accuracy, allowing to estimate ecological/mammalian ADMET properties for novel chemicals. AdmetSAR is accessible free of charge at http://www.admetexp.org.



INTRODUCTION

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug candidates, pesticides, and industrial chemicals play key roles in drug discovery and environmental hazard assessment. Today, drug discovery is a very complex and costly endeavor, which includes disease selection, target identification, lead or hit discovery and optimization, and preclinical and clinical trials.[1] In the past decade, only a few drugs out of hundreds of candidates finally reached the market due to the high failure rate at the clinical trial stage. Two main causes to these failures are the lack of efficacy and unacceptable toxicity. Before 10 years ago, about 50% of potential therapeutic compounds failed in clinical trials or were removed from the market due to unacceptable side-effects and poor ADMET properties.[1] In fact, it is now far less (about 8%) compounds that fail due to poor ADMET properties, which is because these get much more attention now.[2] Filtering and optimization of ADMET properties in the early stage of the drug discovery are intensively investigated.[3] However, the experimental evaluation of ADMET profiles is costly, and the work load can not meet the demands of drug screening and lead optimization. In conjunction with high throughput in vitro screening, computational techniques that can filter/predict ADMET profiles have became an alternative approach.

As of April 2012, there were more than 67 million chemicals registered in the US Chemical Abstracts Service (CAS) database (http://www.cas.org/). Definitely some of these synthetic chemicals improved the quality of our life. However, they have also brought serious environmental pollution and greatly increase our health and ecological risk exposure. For example, a lot of pesticides and industrial chemicals, such as dichlorodiphenyltrichloroethane (DDT), chlordane, and dieldrin, were removed from market due to their environmental persistence, bioaccumulation, and toxicity (PBT) properties.[4] Due to the lack of comprehensive experimental data, high study cost, and animal welfare, the use of computational approaches for assessing the PBT profiles of chemicals is encouraged.[5] Several famous regulatory organizations, such as the Organization for Economic Co-operation and Development (OECD), International Organization for Standardization (ISO), Japanese Ministry of International Trade and Industry (MITI), National Institute of Technology and Evaluation (NITE), European Union (EU), and the United States Environmental Protection Agency (US-EPA) have been developing computational methods and techniques for chemical hazard prediction in environmental risk assessment. It is very urgent to develop new methods, tools, and innovative frameworks, such as a toxicity-associated database and server for chemical safety profiling and environmental hazard assessment.

Received: August 4, 2012
Published: October 23, 2012

   dx.doi.org/10.1021/ci300367a | J. Chem. Inf. Model. 2012, 52, 3099−3105

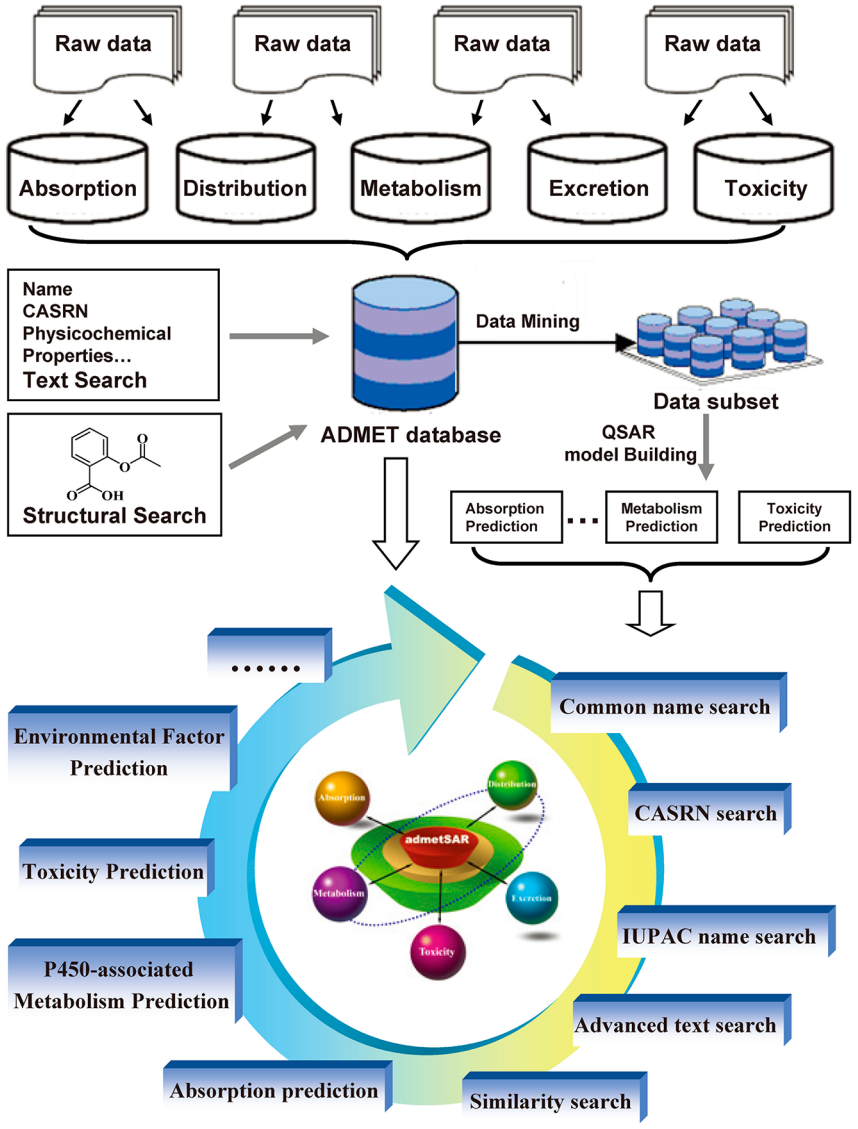


Figure 1. Illustration of pipeline of data curation, data organization, predictive model building, and the application fields of admetSAR.

In the past decade, several specialized ADMET-associated databases and servers, such as FragmentStore,[6] SuperToxic,[7] SuperTarget,[8] FAF-Drugs,[9] T3DB,[10] ADME-AP,[11] PK/DB,[12] PKKB,[13] PreADME (http://www.bmdrc.org/04_product/01_preadme.asp), OCHEM[14] (http://ochem.eu), VCCLAB (http://www.vcclab.org), and CRDD tools (http://crdd.osdd.net/admet.php) were developed for ADMET property filtering. However, most of databases or servers have limits due to narrow chemical space coverage. For example, the PK/PB only collected 1389 compounds with 4141 pharmacokinetic measurement for 8 ADME end points.[12] The recent developed PKKB only incorporates 1685 drugs with about 10 000 experimental ADMET measurements.[13] Recently, several commercially available database and tools, such as ADMET-Predictor (http://www.simulations-plus.com/Default.aspx), ACD/Laboratories' Suite (http://www.acdlabs.com/home/), and FUJITSU ADME Database (http://jp.fujitsu.com/group/kyushu/en/services/admedatabase/index.html) were developed. These databases limit the free use due to expensive prices. Although there are numerous databases developed (mostly proprietary data from the industry from regulatory agencies), high quality and comprehensive ADMET prediction tools are usually not freely available online to general public.[15]

Here, we reported an open source, comprehensive computer readable database, namely admetSAR, to filter or predict ADMET-associated properties of diverse molecules. The original data sets (http://www.lmmd.org/database/cheminformatics/) stored in admetSAR have been used by more than 20 scientific groups from pharmaceutical industries and academics. In admetSAR, over 210 000 high quality ADMET annotated data points for more than 96 000 unique compounds have been carefully curated from a large number of diverse literatures. And, 27 quantitative structure−activity relationship (QSAR) models, including 22 qualitative classification and 5 quantitative regression models developed by ourselves[4,16−20] and other groups,[21−31] were implemented for ADMET property prediction of novel molecules.

METHODS

Database Source. *Data Collection.* The core ADMET-associated data in admetSAR are extracted from the full text of peer-reviewed scientific publications via weekly PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and Google Scholar

dx.doi.org/10.1021/ci300367a | J. Chem. Inf. Model. 2012, 52, 3099−3105

(http://scholar.google.com/) searches from 2002 to 2011. A literature search was performed with general items: "computational (in silico) ADME", "computational toxicology", etc. The ADMET-associated keywords, such as water solubility, human intestinal absorption, oral bioavailability, blood−brain barrier penetration, transporter, plasma protein binding, volume of distribution, CYP450, toxicity, etc., were used to refine the research results. Review articles and the publication with data points less than 10 were removed. In order to control the quality of data points, only publications in a variety high quality journals, such as *The Journal of Chemical Information and Modeling*, *Molecular Informatics*, *Chemical Research in Toxicology*, *Journal of Medicinal Chemistry*, *Bioorganic Medicinal Chemistry Letters*, etc., were selected for further manual checking.

Data Extraction and Preparation. As given in Figure 1, from each eligible publication, detailed raw data of the compounds tested, any ADMET-associated protein, organism, or specie information for these assays are abstracted and manually checked by our experts. Some data points and structure of chemicals were downloaded from the supporting information of publications. The fuzzy, uncertain, obviously uncorrected data points were removed. At last, molecules with compound name, ADMET end points, detailed test materials, structural information, and data source (full journal name) were collected. CAS registry number (CASRN) information of each environmental chemical or drug was extracted from US-EPA ACToR (http://actor.epa.gov/actor/faces/ACToRHome.jsp) and DrugBank databases.[32] Structure of each compound are draw in full by SMILES format and were converted canonical SMILES using the OpenBabel v2.3.1.[33] The IUPAC name of all molecules was generated using Marvin v5.8.0 (http://www.chemaxon.com/). The DrugBank ID number were validated by mapping with DrugBank database v3.0.[32]

Calculation of Physicochemical Property. Calculations of small molecular physicochemical properties are important for computationally filtering their "druglikeness" and "leadlikeness" and toxicity potentials. In admetSAR, five classic physicochemical properties, namely the number of hydrogen bond acceptors and donors, Log P, topological polar surface area (TopoPSA), and molecular weight (MW) were calculated for all compounds using OpenBabel v2.3.1.[33]

Development of Computational Models. Prediction of ADMET-associated properties of new chemicals is a big challenge in free ADMET research communities. In admetSAR, 22 qualitative classification models were implemented, which were developed using support vector machine classification algorithm and in house substructure pattern recognition method.[20] In addition, five quantitative regression models were also built and implemented, which were developed using support vector machine regression algorithm. The robustness of all models was validated based on 5-fold cross validation, and the predictability of several models was validated using available external validation sets. Only models with high predictive accuracy were selected and implemented in admetSAR. In the process of model development, all compounds were represented using MACCS keys implemented with OpenBabel v2.3.1.[33] The detailed descriptions of model building procedure, modeling algorithms, and model validation criteria were given in the Supporting Information.

Database Design and Implementation. The data extracted from publications and manually checked were managed through MySQL v5.1.61 (http://www.mysql.com/). The admetSAR system (Figure 1) was built using Django v1.4.0 on Apache v2.2.20 with mod_wsgi v3.3, installed on Ubuntu-Server v11.10. AdmetSAR provides user-friendly web interfaces to generate a chemical profile, by either text or structural similarity search, and computational prediction using cascading style sheet (CSS) and python script.

Similarity Search. JSDraw v1.3.1 (http://www.scilligence.com/web/download.aspx?prod=JSDraw) was implemented as the build-in molecule editor. The structural similarity search is assessed by the Tanimoto coefficient using the MACCS keys implemented with OpenBabel v2.3.1.[33]

Visualization Features. The two-dimensional (2D) chemical structures were displayed by images, which were generated using Marvin v5.8.0 (http://www.chemaxon.com/).

RESULTS AND DISCUSSION

Description of Database. The admetSAR is accessible free of charge at http://www.admetexp.org. In total, the admetSAR database included more than 210 000 annotated measurements of 95 629 unique compounds (version 1, October 12, 2012), including thousands of FDA approved and experimental drugs, pesticides, environmental agents, and industrial chemicals. The data fields of each compound included three types: the general information, the physicochemical properties, and the ADMET associated profiles. The general data of each molecule includes IUPAC name, formula, CASRN, common name, DrugBank ID, SMILES. The physicochemical properties include MW, Log P, the number of hydrogen bond acceptors and donors, and TopoPSA. More than 45 kinds of ADMET-associated properties, proteins, species, and organisms, such as water solubility, human intestinal absorption, oral bioavailability, blood−brain barrier penetration, P-glycoprotein substrate and inhibitor, renal organic cation transporter, plasma protein binding, volume of distribution, CYP450 substrates and inhibition (CYP1A2, 2C9, 2C19, 2D6 and 3A4), drug-induced liver injury, human Ether-a-go-go-Related gene (hERG) inhibition, rat acute toxicity, skin sensitivity, AMES mutagenicity, carcinogens, fish toxicity, Tetrahymena pyriformis toxicity, honey bee toxicity, quail toxicity, reproductive toxicity, biodegradability, bioconcentration factors, etc., were stored in admetSAR database (Table 1). The detailed biological end points including K_i, IC_{50} (half maximal inhibitory concentration), LC_{50} (median lethal concentration), LD_{50} (median lethal dose), IGC_{50} (50% growth inhibitory concentration), AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control), EC_{50} (half maximal effective concentration), TD_{50} (median toxic dose), etc. were stored in admetSAR.

Prediction of ADMET Properties of New Chemicals. In total, 22 highly predictive qualitative classification models were implemented (version 1, October 12, 2012). These models includes human intestinal absorption, blood-brain barrier penetration, Caco-2 permeability, P-glycoprotein substrate and inhibitor, CYP450 substrate and inhibitor (CYP1A2, 2C9, 2D6, 2C19, and 3A4), hERG inhibitors, AMES mutagenicity, carcinogens, fathead minnow toxicity, honey bee toxicity, and Tetrahymena Pyriformis toxicity (Table 1). The range of the area under the receiver operating characteristic curve (AUC) is from 0.638 to 0.956 for 22 classification models (Supporting Information Table S1) via 5-fold cross validation. In addition, all classification models were given a probability output described in our previous work,[17] instead of simple binary output. In scientific community of ADMET prediction, quantitative predictions are more useful. Therefore,

Table 1. Important Data Fields, the Number of High-Throughput Screening (HTS) or NonHTS Data Measurements, Whether the Data Was Used to Build a Model and References to the Original Sources in AdmetSAR

no.	end points	number of measurements	HTS data yes/no	model building yes/no	ref
1	aqueous solubility (I)	1708	no	yes	36
2	aqueous solubility (II)	46315	yes	no	37
3	human intestinal absorption	578	no	yes	20
4	Caco-2 permeability	674	no	yes	29
5	blood–brain barrier	1839	no	yes	20
6	P-gp substrate	332	no	yes	31
7	P-gp inhibitor (I)	1273	no	yes	23
8	P-gp inhibitor (II)	1275	yes	yes	21
9	CYP1A2 inhibitor	14903	yes	yes	17
10	CYP2C9 inhibitor	14709	yes	yes	17
11	CYP2C19 inhibitor	14576	yes	yes	17
12	CYP2D6 inhibitor	14741	yes	yes	17
13	CYP3A4 inhibitor	18561	yes	yes	17
14	CYP2C9 substrate	673	no	yes	22
15	CYP2D6 substrates	671	no	yes	22
16	CYP3A4 substrates	671	no	yes	22
17	hERG inhibitor (I)	368	no	yes	28
18	hERG inhibitor (II)	806	no	yes	30
19	AMES mutagenicity	8445	no	yes	38
20	chemical carcinogens	293	no	yes	27
21	fathead minnow toxicity	554	no	yes	19
22	honey bee toxicity	195	no	yes	19
23	tetrahymena pyriformis toxicity	1571	no	yes	16
24	rat acute toxicity	10207	no	yes	39
25	hepatotoxicity	2154	no	no	40,41
26	reproductive toxicity	4621	yes	no	42
27	maximum recommended daily dose	1214	no	no	43
28	biodegradation	947	no	yes	4
29	bioconcentration factors	916	no	no	44

five highly predictive quantitative regression models including Caco-2 permeability, water solubility, rat acute toxicity, Tetrahymena pyriformis toxicity, and fathead minnow acute toxicity were built using the support vector machine (SVM) regression algorithm and implemented in admetSAR. The range of the square of correlation coefficient (R^2) is from 0.564 to 0.810, and the range of root-mean-square deviation (RMSE) is from 0.256 to 0.283 using 5-fold cross validation (Supporting Information Table S2). To ensure usefulness of admetSAR, it will be updated monthly with additional computational models based on available data, especially for quantitative models using MySQL database management system. If data sets with new end points are reported, new models will be built and implemented in our database. If data sets with known end points in admetSAR are reported, the old models will be updated and replaced by new ones with more diverse chemical space coverage.

The generalization ability of a model decides the usefulness and reliability of models. In order to test the actual predictive ability of admetSAR, several models were validated using the available external validation sets. For example, 27 novel chemicals were predicted first using the admetSAR server and were further assayed using the MITI-I test protocol.[4] The detailed experimental and predicted results were given in Supporting Information Scheme S1 and Table S3. The overall predictive accuracy of admetSAR was 88.9%; that is, 24 chemicals were predicted correctly. The admetSAR outperformed Biowin5 and Biowin6 implmented in the EPI Suite v4.10 (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm). The detailed results of the external validation sets and the comparison with published models can be found in Supporting Information Tables S4 and S5.

In addition, above data sets of 27 computational models were released. All data sets were carefully curated by us and can be directly used to build ADMET models and methodology assessment, which can be downloaded form the download module (http://www.admetexp.org/download/). However, a registration is required for all users. And we encourage the user to provide new comments about how to improve the next version of admetSAR or upload new data sets to admetSAR.

Interface and Data Management. In admetSAR, user-friendly, web-based query tools incorporating a molecular build-in interface enables the database to be queried by common name, CASRN, IUPAC name, SMILES, and structural similarity search (Figure 2). The user can use an advanced search as a useful way for database querying. Here, admetSAR provides three different levels of chemical ADMET properties filtering (Figures 1 and 2). In first level, if a molecule had been retrieved in the admetSAR database, the user can find useful information of this molecule using the basic search or advanced search modules. In the second level, if a molecule was not retrieved in database, the user can query potential useful information of its analogue using the structural similarity search. In the last level, if a molecule was not retrieved and there is no high similar molecule in the database, the user could predict the novel ADMET properties using the 27 predictive QSAR models by simple SMILES input or molecule build-in tool of JSDraw. Therefore, admetSAR could provide a comprehensive prediction or filtering of ADMET profiles based on different criteria of users.

SUMMARY AND PERSPECTIVES

In summary, we described here a large database which provides a comprehensive estimation of key ADMET profiles for diverse compounds. In admetSAR, over 210 000 ADMET annotated data points for more than 96 000 unique compounds with 45 kinds of ADMET-associated properties, proteins, species, or organisms have been carefully curated from a large amount of diverse literatures. Moreover, 22 predictive qualitative classification models with probability output and 5 quantitative regression models with actual number output were developed and implemented for predicting ADMET properties of novel diverse molecules. Therefore, admetSAR will facilitate researchers to freely predict ADMET properties and develop more useful ADMET predictive models in the further.

We are continuing to improve admetSAR in the following directions. First, to ensure its usefulness, the database will be updated monthly or quarterly with additional published data

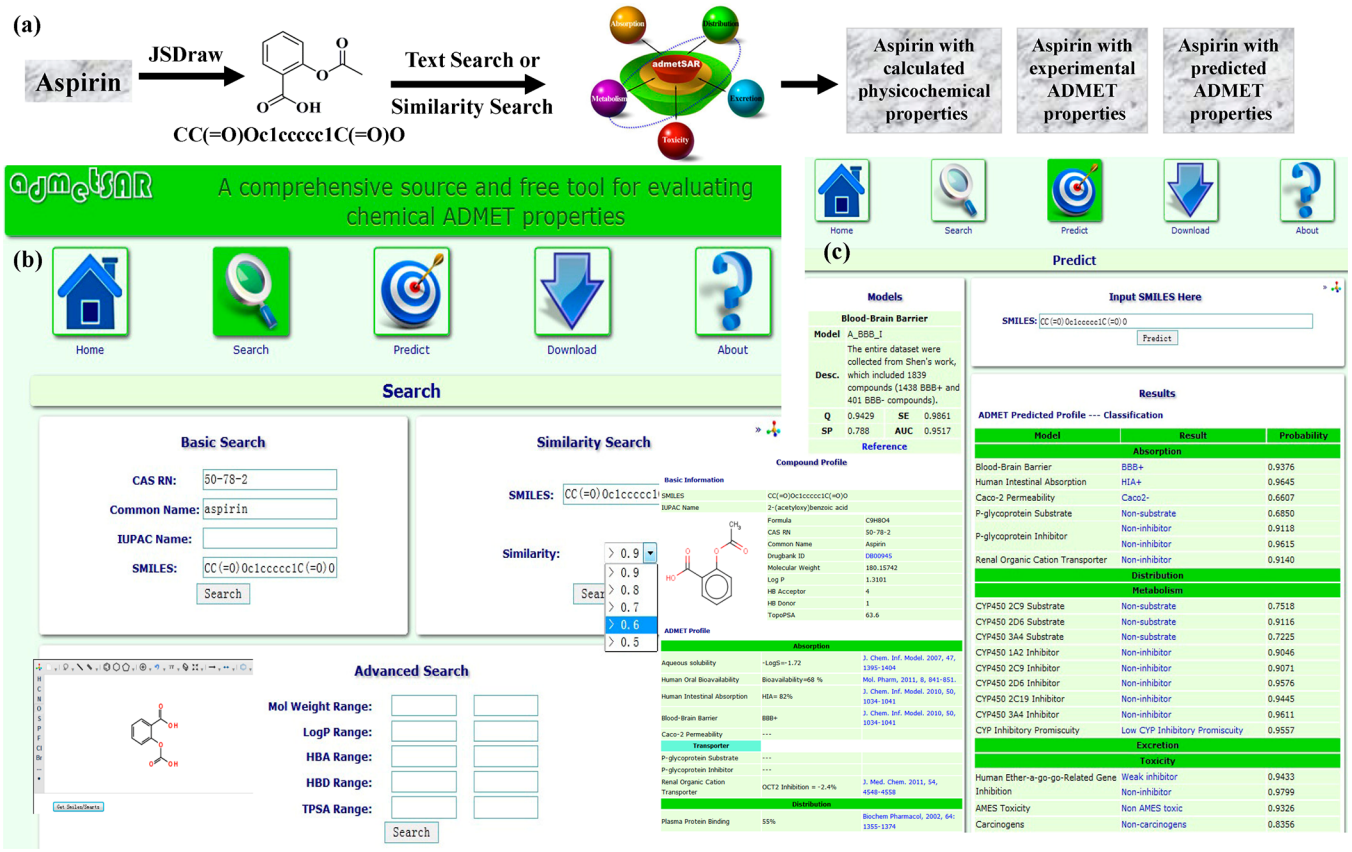


Figure 2. AdmetSAR allowing searching or predicting chemical ADMET profiles using a user-friendly interface. (a) Overview of the admetSAR Pipeline. (b) Search modules: basic search, advanced search, and structural similarity search. (c) Results of predicting ADMET properties of novel molecules using the SMILES input or build-in molecule editor of JSDraw by predictive computational models.

and models using our MySQL database management systems, especially high quality of toxicity associated data sets and quantitative regression models for predicting ADMET properties of novel molecules. Although admetSAR already provided the comprehensive quality data and computational models for a large number of ADMET-associated end points, some data fields are missing due to data not being available. Second, the quality of manually curated data should be improved further. For example, we searched the literature about the bioavailability of aspirin and found that there were several different results. For example, Woodford and Lesko reported that the relative bioavailability of aspirin gum was 69.5 ± 3.4%, based on cumulative 24 h urinary excretion of total salicylate after the chewing of three gum tablets for 15 min.[34] Pedersen et al. reported that systemic bioavailability ranged from 46 to 51% of single oral doses of 20, 40, 325, and 1300 mg of aspirin on five healthy volunteers.[35] Recently, Tina et al. used the bioavailability of aspirin 68% for building computational oral bioavailability models. Here, the more reliable bioavailability data of aspirin 68% was stored in our database. In next version of admetSAR, the drug with more additional data, such as formulation, dose-levels, testing condition and period, etc., will be added and updated. In addition, more toxicity-associated source, such as phenotypic information of drug adverse events (http://www.lmmd.org/online_services/metaadedb/), will be added in admetSAR for comprehensively ADMET properties filtering and personalized medicine.

■ ASSOCIATED CONTENT

Supporting Information

Detailed descriptions of model building procedure, modeling algorithms, and model validation criteria, Scheme S1, Tables S1−S5. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +86-21-6425-1052. Fax: +86-21-6425-3651. E-mail: ytang234@ecust.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the 863 Project (Grant 2012AA020308), the National Natural Science Foundation of China (Grant 21072059), the Fundamental Research Funds for the Central Universities (Grant WY1113007, WY1014010), the National S&T Major Project of China (Grant 2011ZX09307-002-03), and the Shanghai Committee of Science and Technology (Grant 11DZ2260600).

■ REFERENCES

(1) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug. Discovery* **2004**, *3*, 711−715.
(2) Merlot, C. Computational toxicology−a tool for early safety evaluation. *Drug Discovery Today* **2010**, *15*, 16−22.

(3) Hou, T.; Wang, J. Structure-ADME relationship: still a long way to go? *Expert. Opin. Drug Metab. Toxicol.* **2008**, *4*, 759−770.
(4) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In Silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655−669.
(5) Rusyn, I.; Daston, G. P. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ. Health Perspect.* **2010**, *118*, 1047−1050.
(6) Ahmed, J.; Worth, C. L.; Thaben, P.; Matzig, C.; Blasse, C.; Dunkel, M.; Preissner, R. FragmentStore–a comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic Acids Res.* **2011**, *39*, D1049−1054.
(7) Schmidt, U.; Struck, S.; Gruening, B.; Hossbach, J.; Jaeger, I. S.; Parol, R.; Lindequist, U.; Teuscher, E.; Preissner, R. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.* **2009**, *37*, D295−299.
(8) Hecker, N.; Ahmed, J.; von Eichborn, J.; Dunkel, M.; Macha, K.; Eckert, A.; Gilson, M. K.; Bourne, P. E.; Preissner, R. SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.* **2012**, *40*, D1113−1117.
(9) Miteva, M. A.; Violas, S.; Montes, M.; Gomez, D.; Tuffery, P.; Villoutreix, B. O. FAF-Drugs: free ADME/tox filtering of compound collections. *Nucleic Acids Res.* **2006**, *34*, W738−744.
(10) Lim, E.; Pon, A.; Djoumbou, Y.; Knox, C.; Shrivastava, S.; Guo, A. C.; Neveu, V.; Wishart, D. S. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic Acids Res.* **2010**, *38*, D781−786.
(11) Sun, L. Z.; Ji, Z. L.; Chen, X.; Wang, J. F.; Chen, Y. Z. ADME-AP: a database of ADME associated proteins. *Bioinformatics* **2002**, *18*, 1699−1700.
(12) Moda, T. L.; Torres, L. G.; Carrara, A. E.; Andricopulo, A. D. PK/DB: database for pharmacokinetic properties and predictive computational ADME models. *Bioinformatics* **2008**, *24*, 2270−2271.
(13) 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. *J. Chem. Inf. Model.* **2012**, *52*, 1132−1137.
(14) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 533−554.
(15) van de Waterbeemd, H.; Gifford, E. ADMET computational modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(16) Cheng, F.; Shen, J.; Yu, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* **2011**, *82*, 1636−1643.
(17) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and non-Inhibitors using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996−1011.
(18) Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P. W.; Tang, Y. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J. Chem. Inf. Model.* **2011**, *51*, 2482−2495.
(19) Cheng., F.; Shen, J.; Li, W.; Lee, P. W.; Tang, Y. In Silico prediction of terrestrial and aquatic toxicities for organic chemicals. *Chin. J. Pestic. Sci.* **2010**, *12*, 477−488.
(20) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034−1041.
(21) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.* **2011**, *54*, 1740−1751.
(22) Carbon-Mangels, M.; Hutter, M. C. Selecting relevant descriptors for classification by bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets. *Mol. Inf.* **2011**, *30*, 885−895.
(23) Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharmaceutics* **2011**, *8*, 889−900.
(24) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Muller, K. R. Benchmark data set for computational prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009**, *49*, 2077−2081.
(25) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of non-congeneric compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.
(26) Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006**, *46*, 597−605.
(27) Lagunin, A.; Filimonov, D.; Zakharov, A.; Xie, W.; Huang, Y.; Zhu, F.; Shen, T.; Yao, J.; Poroikov, V. Computer-aided prediction of rodent carcinogenicity by PASS and CISOC-PSCT. *QSAR Comb. Sci.* **2009**, *28*, 806−810.
(28) Robinson, R. M.; Glen, R. C.; Mitchell, J. B. Development and comparison of hERG blocker classifiers: assessment on different datasets yields markedly different results. *Mol. Inf.* **2011**, *30*, 443−458.
(29) The, H. P.; Gonzalez Alvarez, I.; Bermejo, M.; Sanjuan, V. M.; Centelles, I.; Garrogues, T. M.; Cabrera Perez, M. A. Computational prediction of Caco-2 cell permeability by a classification QSAR approach. *Mol. Inf.* **2011**, *30*, 376−385.
(30) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharmaceutics* **2012**, *9*, 996−1010.
(31) Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R. C.; Yan, A. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.* **2011**, *51*, 1447−1456.
(32) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−1041.
(33) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(34) Woodford, D. W.; Lesko, L. J. Relative bioavailability of aspirin gum. *J. Pharm. Sci.* **1981**, *70*, 1341−1343.
(35) Pedersen, A. K.; FitzGerald, G. A. Dose-related kinetics of aspirin. Presystemic acetylation of platelet cyclooxygenase. *N. Engl. J. Med.* **1984**, *311*, 1206−1211.
(36) Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395−1404.
(37) Cheng, T.; Li, Q.; Wang, Y.; Bryant, S. H. Binary classification of aqueous solubility using support vector machines with reduction and recombination feature selection. *J. Chem. Inf. Model.* **2011**, *51*, 229−236.
(38) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico prediction of chemical ames mutagenicity. *J. Chem. Inf. Model.* **2012**, DOI: 10.1021/ci300400a.
(39) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure-activity relationship modeling of rat

acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.
(40) 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*, 171−183.
(41) Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem. Res. Toxicol.* **2010**, *23*, 724−732.
(42) Huang, R.; Xia, M.; Cho, M. H.; Sakamuru, S.; Shinn, P.; Houck, K. A.; Dix, D. J.; Judson, R. S.; Witt, K. L.; Kavlock, R. J.; Tice, R. R.; Austin, C. P. Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. *Environ. Health Perspect.* **2011**, *119*, 1142−1148.
(43) 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. *Curr. Drug Discovery Technol.* **2004**, *1*, 61−76.
(44) Zhao, C.; Boriani, E.; Chana, A.; Roncaglioni, A.; Benfenati, E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* **2008**, *73*, 1701−1707.