JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

ADMET Evaluation in Drug Discovery. 11. PharmacoKinetics Knowledge Base (PKKB): A Comprehensive Database of Pharmacokinetic and Toxic Properties for Drugs

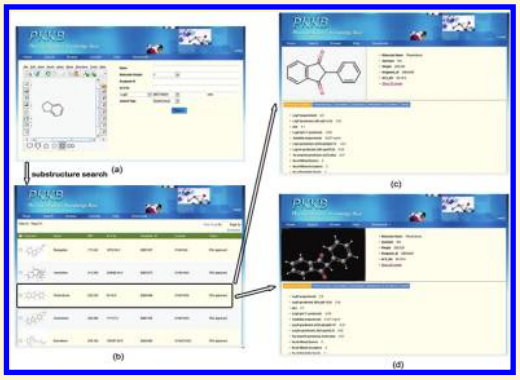
Dongyue Cao,[†,∥] Junmei Wang,[§,∥] Rui Zhou,[†] Youyong Li,[†] Huidong Yu,[†] and Tingjun Hou*[,†,‡]

[†]Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China

[‡]College of Pharmaceutical Science, Soochow University, Suzhou, Jiangsu 215123, China

[§]Department of Biochemistry, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, United States

ABSTRACT: Good and extensive experimental ADMET (absorption, distribution, metabolism, excretion, and toxicity) data is critical for developing reliable in silico ADMET models. Here we develop a **P**harmaco**K**inetics **K**nowledge **B**ase (PKKB) to compile comprehensive information about ADMET properties into a single electronic repository. We incorporate more than 10 000 experimental ADMET measurements of 1685 drugs into the PKKB. The ADMET properties in the PKKB include octanol/water partition coefficient, solubility, dissociation constant, intestinal absorption, Caco-2 permeability, human bioavailability, plasma protein binding, blood-plasma partitioning ratio, volume of distribution, metabolism, half-life, excretion, urinary excretion, clearance, toxicity, half lethal dose in rat or mouse, etc. The PKKB provides the most extensive collection of freely available data for ADMET properties up to date. All these ADMET properties, as well as the pharmacological information and the calculated physiochemical properties are integrated into a web-based information system. Eleven separated data sets for octanol/water partition coefficient, solubility, blood-brain partitioning, intestinal absorption, Caco-2 permeability, human oral bioavailability, and P-glycoprotein inhibitors have been provided for free download and can be used directly for ADMET modeling. The PKKB is available online at http://cadd.suda.edu.cn/admet.



INTRODUCTION

Drug discovery and development is a time-consuming and expensive process. It was estimated that 40−60% of new chemical entity (NCE) failures can be attributed to poor ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiles.[1,2] ADMET properties can be predicted from the chemical structures, so that huge number of compounds can be evaluated prior to be synthesized and assayed.[3−5] Theoretical predictions of ADMET properties have been proven to be efficient in recent years.[3−7] The lack of enough high quality experimental data for training reliable models has been the major hurdle to model ADMET properties.[6,8,9] When the sample size used in training is limited, the in silico models cannot give robust and accurate predictions, especially for the ADMET properties involving complex processes, such as bioavailability, metabolism, toxicity, etc.[6] Traditionally, the available experimental data sets for ADMET modeling in the public domain are often limited in quantity and quality. This is particularly true for in vivo properties obtained directly from human, where data is typically only available for compounds in clinic development.[8] Encouragingly, the available large data sets are expanding in the recent years. For example, three extensive data sets for intestinal absorption, oral bioavailability in human, and P-gp inhibitors were reported by our group.[10−13] Nevertheless, further developments on the availability of ADMET data for the public use are still necessary.

With more available ADMET data, it will be helpful to integrate all these data of a variety of ADMET properties from different sources into a single information system. The PK/DB reported by Moda and the co-worker is one information system providing the service.[14] The PK/DB incorporates 1389 compounds and 4141 pharmacokinetic measurements for 8 ADME properties. And the data in the PK/DB were directly taken from the reported publicly available ADME data sets without careful curation. For instance, in the PK/DB, two core data sets, the intestinal absorption data set with 687 molecules and the oral bioavailability data set with 660 molecules, were directly taken from the data sets reported by us.[11,12] Thus the PK/DB only provides a limited data information.

Here, we develop the PKKB (**P**harmaco**K**inetics **K**nowledge **B**ase) to house 2- and 3D chemical structures, pharmacological information, experimental or calculated physiochemical properties, and particularly high quality ADMET data of drugs.

Received: February 29, 2012

   dx.doi.org/10.1021/ci300112j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

The PKKB incorporates the most extensive collections of ADMET data in the public domain, which is much larger than that in the PK/DB. The total number of experimental data in the PKKB is more than 10 000, in comparison with 4000 in the PK/DB. Most data in the PKKB were collected by us to develop the ADMET prediction models. During the ADMET modeling process, the experimental data were carefully curated by us.[10−13,15−20] The data sets in the PKKB developed by us were widely used by a lot of well-known information systems, such as Drugbank,[21] KnowItAll,[22] and PK/DB,[14] and scientists from pharmaceuticals and academics. The extensive feedbacks from the users are very helpful for the improvement of the data in the PKKB. We frequently update the data in the PKKB, which is important to improve the quality of data. As a publicly available online database, the PKKB will be continuously maintained and updated.

METHODOLOGY

Content of the Database. The PKKB is hosted at http://cadd.suda.edu.cn/admet. The ADMET data currently in the PKKB are from 1685 drugs. All the FDA-approved small-molecule drugs found in DrugBank have been collected into the PKKB.[21] We categorize the data fields for each molecule into four groups: the general information, the pharmacological information, the physicochemical properties, and the ADMET properties (Table 1). The general information for each molecule includes the molecular name, synonyms, ACS number, and DrugBank ID. The pharmacological information for each molecule includes the status, the administration route, and the pharmacological effect. The physiochemical properties for each molecule include the experimental octanol/water partition coefficient (log P), the experimental aqueous solubility (log S), the experimental dissociation constant (pK_a), the calculated octanol/water partition coefficient (log P), the calculated octanol/water distribution coefficient at pH = 7 (log D), the calculated aqueous solubility, the number of hydrogen bond donors, the number of hydrogen bond acceptors, the number of rotatable bonds, and the topological polar surface area (TPSA). We performed all the calculations with ACD/Laboratories (version 12.0). The ADMET properties for each molecule in the PKKB are categorized into five parts: absorption, distribution, metabolism, excretion, and toxicity. The properties associated with absorption include the absolute value of intestinal absorption, the description of intestinal absorption, Caco-2 permeability, and bioavailability in human. The properties associated with distribution include protein binding, volume of distribution (VD) and blood/plasma partitioning ratio (D-blood). The properties associated with metabolism include general metabolism information and half-time; the properties associated with distribution include excretion route, urinary excretion, and clearance. The properties associated with toxicity include general toxicity information, LD_{50} in rat and LD_{50} in mouse. The distributions for ten ADMET properties are shown in Figure 1.

We also release eleven ADMET data sets reported by us in the PKKB. These ADME data sets include three solubility data sets of 1290, 1708, and 1210 molecules, respectively,[16,23,24] a Caco-2 permeability data set of 100 molecules,[20] a blood−brain partitioning data set of 109 molecules,[18] a P-gp inhibitor data set of 1302 molecules,[10] a intestinal absorption data set of 647 molecules,[11,15] a bioavailability data set of 1013 molecules,[12,13] a hERG blocker data set 806 molecules,[25] a combined data set of 470 compounds with both intestinal absorption data and oral bioavailability data, and a combined data set of 69 compounds with both intestinal absorption data and Caco-2 permeability data.[11−13,20] All these data sets have been postprocessed and can be directly used for ADMET modeling. In the near future, more data sets will be continuously added to the data set collections in the PKKB. These data sets are important for those who are interested in benchmarking the results of experiments, validating the accuracy of existing ADMET predictive models, and building new predictive models.

It must be noted that the purpose of PKKB is quite different from those of CheMBL,[26] ChemSpider,[27] and DrugBank.[21] ChemSpider is a free chemical structure database with more than 25 million molecules, and CheMBL is a free chemical database of bioactive drug-like small molecules. Drugbank is a popular information system for drugs and was primarily developed to provide chemical structures, pharmacological data and drug targets for drugs. ChemSpider does not afford valuable information about ADMET. Some ADMET data are included in DrugBank and ChEMBL, but they are quite limited.

Table 1. Important Data Fields in the PKKB and the Corresponding Number of Measures

no.	property	measures
physiochemical properties		
1	molecular weight	1684
2	log P (experiment)	1019
3	log P (predicted, AB/logP v2.0)	1625
4	pK_a (experiment)	638
5	log D (pH = 7, predicted)	1625
6	solubility (experiment)	800
7	log S (predicted, ACD/Laboratories; pH = 7)	1614
8	log S_w (predicted, AB/LogSw 2.0)	1625
9	S_w (mg/mL) (predicted, ACD/Laboratories)	1613
10	S_w (predicted)	1625
11	number of hydrogen bond donors	1625
12	number of hydrogen bond acceptors	1625
13	Number of rotatable bonds	1625
14	TPSA	1625
pharmacology		
15	status	1372
16	administration	501
17	pharmacology	1543
absorption		
18	intestinal absorption	679
19	absorption (description)	699
20	Caco-2 permeability	64
21	human bioavailability	992
distribution		
22	plasma protein binding	1058
23	volume of distribution (Vd)	646
24	blood/plasma partitioning ratio (D-blood)	66
metabolism		
25	metabolism	1111
26	half-time	1116
excretion		
27	excretion	855
28	urinary excretion	281
29	clearance	410
toxicity		
30	description of toxicity	873
31	LD_{50} (rat)	219
32	LD_{50} (mouse)	243

dx.doi.org/10.1021/ci300112j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

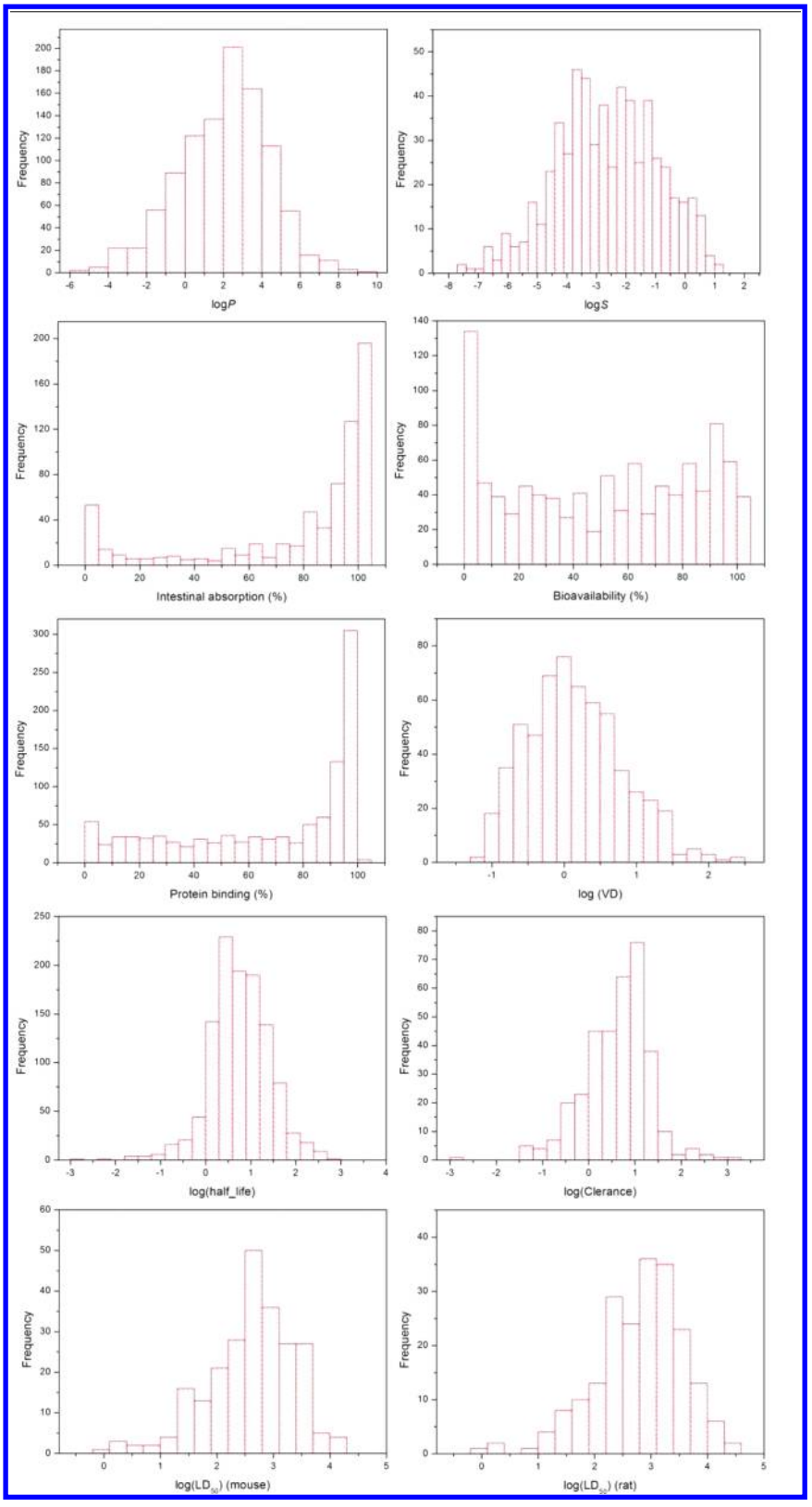


Figure 1. Distributions of ten experimental physiochemical and ADMET properties in the PKKB.

In comparison, the PKKB provides the comprehensive data for ADMET modeling.

Implementation. We developed an integrated data structure and a variety of querying functions to allow easy and efficient retrieval of ADMET data (Figure 2). The PKKB is installed on Windows server workstations. MySQL5.1.46 is used as the relational database management system (RDBMS). An Apache Tomcat 6.0.26 server is used as the web server platform. Meanwhile, we use J2EE (Spring3.0.5+Hibernate3.6.3), jQuery, and DHTML as the web interface. In order to construct

dx.doi.org/10.1021/ci300112j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

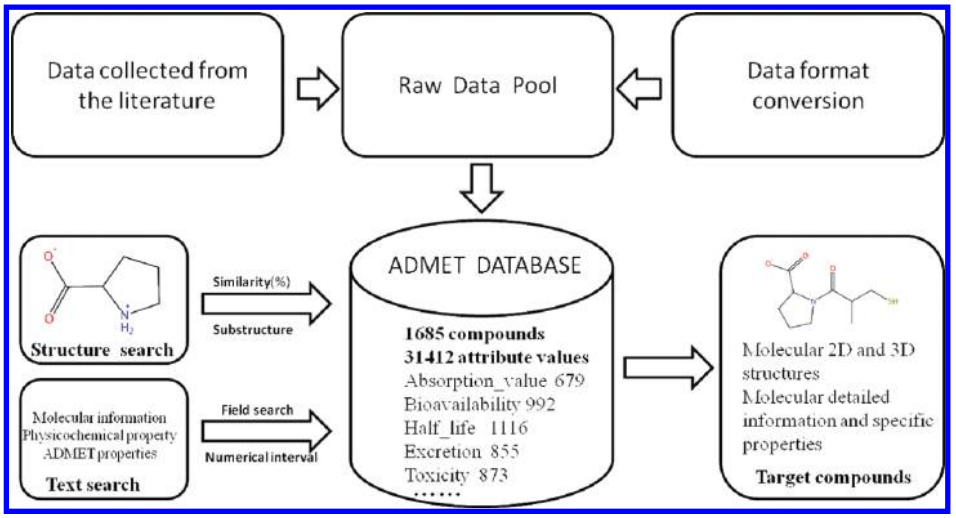


Figure 2. Basic schema of the PKKB. We mine, clean, and organize the ADMET data from the publications by manual curation.

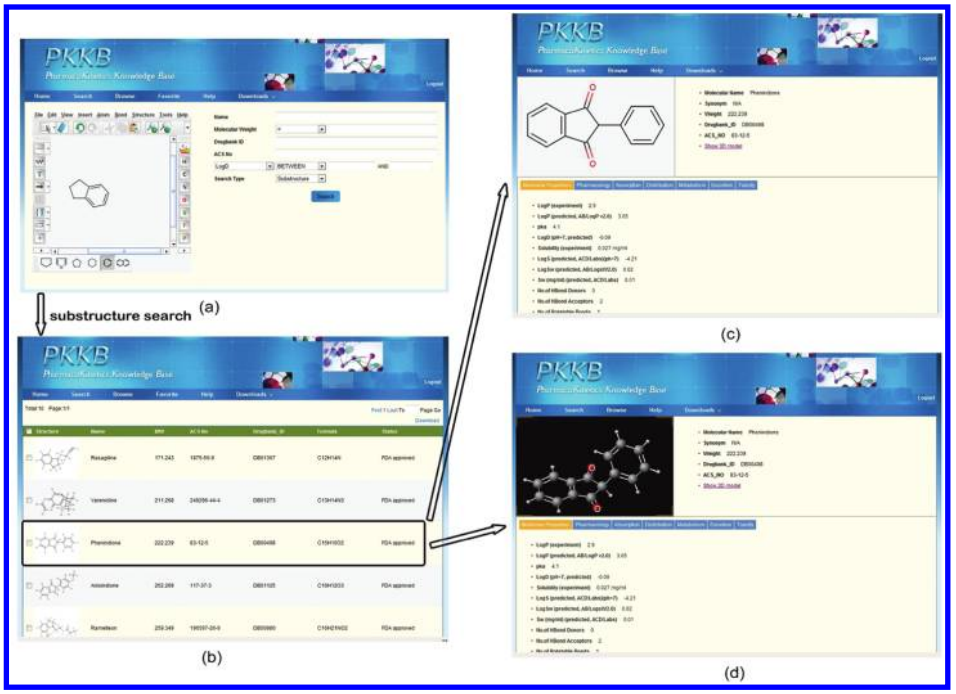


Figure 3. (a) Integrated text and structure searching interface in the PKKB, (b) searching results for a list of target molecules, (c) detailed information of a molecule with 2D structure, and (d) detailed information of a molecule with 3D structure.

user-friendly searching and retrieval systems, the PKKB provides interactive web interfaces based on the graphic structure editor MarvinSketch and the substructure matching algorithm accomplished in OpenBabel2.2.3.

Database Access and Database Query. All querying functions in the PKKB are available for everyone with internet access. However, registration is highly encouraged for all users because the registered users can download the search results using the download hyperlinks that the nonregistered users cannot use. Downloadable results are exported into a structure data format (SDF) database for maximum flexibility of use in other applications. Everyone can download the ten separated ADMET data sets.

The PKKB provides registered users with an integrated searching interface for both structure and text search (Figure 3a). The results from such searches lead the users to a "Molecule list" view (Figure 3b), which shows the basic information about each molecule, including 2D structure, name, molecular weight, DrugBank ID, and ACS number. Users can click the hyperlink of a particular molecule to get more detailed information: 2- and 3D structures, molecular properties, and ADMET properties, which are displayed on its respective information webpage (Figures 3c and d).

The PKKB is incorporated with a web-based query tool supported by MarvinSketch and Openbabel2.2.3. The molecular drawing interface, MarvinSketch, allows users to quickly draw molecules through some basic functions by the graphical user interface (GUI) and advanced functionalities, including sprout drawing, customizable shortcuts, abbreviated groups, default and user defined templates, and context sensitive popup menus. SMARTS rules allow users to define any specific or generic queries. Structural searches contain exact, substructure,

dx.doi.org/10.1021/ci300112j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Figure 4. (a) Compound management interface and (b) the interface for editing the available properties of a specific molecule.

and similarity search supported by Openbabel2.2.3. The similarity between the query and each molecule was measured by Tanimoto coefficient based on the FP2 fingerprint. Users can also take a structural search by inputting a molecule with different molecular formats supported by MarvinSketch.

The PKKB is developed with rapid text searching functions for molecular name, DrugBank ID, and ACS number. Moreover, for several molecular properties, including molecular weight, predicted log *D*, experimental log *P*, predicted log *P*, predicted solubility, number of hydrogen-bond acceptor, number of hydrogen-bond donors, number of rotatable bonds, intestinal absorption, and bioavailability, the numerical interval of these attribute values can be searched to find the target molecules.

Database Management. The administration user of the PKKB can activate the database management system. The database management includes three management interfaces: user management, role management, and compound management. The user management interface is used to manage registered user accounts, such as organizing the user's name, email, phone number, etc. The role management interface is used to setup users' permissions. The compound management interface is used to add or remove a compound in the PKKB (Figure 4a). In the compound management interface, an administrator can edit the available properties of each molecule when it is necessary (Figure 4b); furthermore, the structure of a compound can be edited by a molecular drawing interface supported by MarvinSketch (Figure 4b). In the compound management interface, administrators can take a text search to find a specific compound to be edited (Figure 4a). These functions are helpful for the continuous maintenance of the PKKB.

■ CONCLUSIONS AND FUTURE DEVELOPMENT

In summary, we developed the PKKB, which is a unique knowledge environment for ADMET properties. The PKKB provides structures, pharmacological information, important experimental or predicted physiochemical properties, and experimental ADMET data for 1685 drugs. The PKKB integrates both predicted and experimental information into a single and public resource. We expect that the rich content in the PKKB will facilitate the researchers to develop more reliable ADMET prediction models in the near future. With the extensive data in the PKKB, it is plausible to develop more complicated models for more "complex" ADMET properties, such as bioavailability, clearance, metabolism, etc. For example, from the combined data set of 470 compounds with both intestinal absorption data and oral bioavailability data, we may develop rules or models for the first-pass metabolism effect.

We are continuing to improve the PKKB in the following directions. First of all, the quality of the collected data needs to be improved further. The reliable data can usually be generated under a single experimental protocol or even a single experimental assay. Unfortunately, the data in the PKKB were put together from various sources, and they are subject to variability due to experimental conditions and interlaboratory errors. We will check the reliability of the data from different sources carefully and guarantee the reliability of the data in the PKKB as best as we can. Second, to better characterize each entry, a complete record of that entry is needed, i.e., all the fields for each molecule need to be filled in. Although the PKKB already affords extensive data for many ADMET properties, some data fields are far from "completeness". The empty data field usually indicates that the data has not been measured or reported. However, in many cases, the data probably exists in somewhere else, but our ADMET team has not found it and validated it. Moreover, the coverage of the ADMET properties in the PKKB is still limited. In the new version of the PKKB, some important ADMET properties will be added, such as genotoxicity, aquatic toxicity, eye irritation, skin irritation, P450 inhibitors, and substrates, etc. In addition to the existence of missing data, the PKKB is also missing some drug-like molecules. Currently more than 1000 drug candidates in clinical trials are still on the PKKB "to do" list and will be added in a short period. Third, we will replace Openbabel2.2.3 by MORT (Molecular Objects and Relevant Templates) developed in our group soon. MORT, as

dx.doi.org/10.1021/ci300112j | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the foundation library of *gleap* in AmberTools,[28] has already been released. The Java MORT under development will be used by PKKB. Finally, we plan to afford online prediction models for several important ADMET properties in the next version of PKKB.

AUTHOR INFORMATION

Corresponding Author

*E-mail: tjhou@suda.edu.cn or tingjunhou@hotmail.com.

Author Contributions

‖Co-first authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project is smupported by the National Science Foundation of China (Grant No. 20973121), the National Basic Research Program of China (973 program, 2012CB932600 to T.H.), the NIH (R21GM097617 to J.W.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

REFERENCES

(1) Kennedy, T. Managing the drug discovery/development interface. *Drug Discovery Today* **1997**, *2*, 436−444.

(2) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2004**, *3*, 711−715.

(3) Hou, T. J. In Silico Predictions of ADME/T Properties: Progress and Challenges. *Combin. Chem. High Throughput Screen.* **2011**, *14*, 306−306.

(4) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653−2667.

(5) Zhu, J. Y.; Wang, J. M.; Yu, H. D.; Li, Y. Y.; Hou, T. J. Recent Developments of In Silico Predictions of Oral Bioavailability. *Combin. Chem. High Throughput Screen.* **2011**, *14*, 362−374.

(6) Hou, T.; Wang, J. Structure - ADME relationship: still a long way to go? *Exp. Opin. Drug Metabol. Toxicol.* **2008**, *4*, 759−770.

(7) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(8) Gola, J.; Obrezanova, O.; Champness, E.; Segall, M. ADMET property prediction: The state of the art and current challenges. *Qsar Combin. Sci.* **2006**, *25*, 1172−1180.

(9) Dearden, J. C. In silico prediction of ADMET properties How far have we come? *Exp. Opin. Drug Metabol. Toxicol.* **2007**, *3*, 635−639.

(10) Chen, L.; Li, Y. Y.; Zhao, Q.; Peng, H.; Hou, T. J. 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.

(11) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208−218.

(12) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *J. Chem. Inf. Model.* **2007**, *47*, 460−463.

(13) Tian, S.; Li, Y. Y.; Wang, J. M.; Zhang, J.; Hou, T. J. ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints. *Mol. Pharmaceutics* **2011**, *8*, 841−851.

(14) Moda, T. L.; Torres, L. G.; Carrara, A. E.; Andricopulo, A. D. PK/DB: database for pharmacokinetic properties and predictive in silico ADME models. *Bioinformatics* **2008**, *24*, 2270−2271.

(15) Hou, T. J.; Wang, J. M.; Li, Y. Y. ADME evaluation in drug discovery. 8. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408−2415.

(16) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

(17) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery - 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *J. Mol. Model.* **2002**, *8*, 337−349.

(18) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors (vol 43, 2137, 2003). *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 766−770.

(19) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible surface areas. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1058−1067.

(20) Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1585−1600.

(21) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.

(22) *KnowItAll information system*; Bio-Rad Laboratories, Inc. http://www.knowitall.com, 2012.

(23) Wang, J. M.; Hou, T. J.; Xu, X. J. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571−581.

(24) Wang, J. M.; Krudy, G.; Hou, T. J.; Zhang, W.; Holland, G.; Xu, X. J. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395−1404.

(25) 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.

(26) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(27) Pence, H. E.; Williams, A. ChemSpider: an online chemical information resource. *J. Chem. Educ.* **2010**, *87*, 1123−1124.

(28) Zhang, W.; Hou, T. J.; Qiao, X. B.; Xu, X. J. Some basic data structures and algorithms for chemical generic programming. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1571−1575.