Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

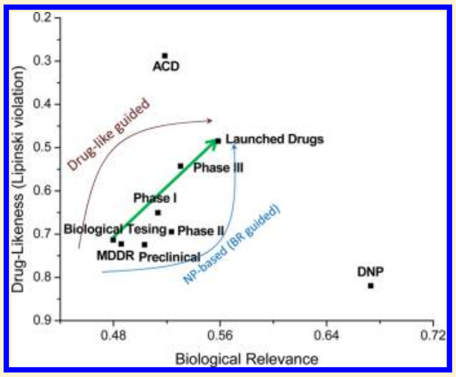
Exploring the Biologically Relevant Chemical Space for Drug Discovery

Zhi-Luo Deng,[†] Cai-Xia Du,[‡] Xiao Li,[∥,⊥] Ben Hu,[†] Zheng-Kun Kuang,[†] Rong Wang,[‡] Shi-Yu Feng,[‡] Hong-Yu Zhang,[‡,§] and De-Xin Kong*,[†,‡]

[†]State Key Laboratory of Agricultural Microbiology, [‡]Center for Bioinformatics, College of Life Science and Technology, and [§]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

[∥]Center for Bioinformatics, College of Life Science, Shandong University of Technology, Zibo 255049, China

Ⓢ *Supporting Information*



ABSTRACT: Both recent studies and our calculation suggest that the physicochemical properties of launched drugs changed continuously over the past decades. Besides shifting of commonly used properties, the average biological relevance (BR) and similarity to natural products (NPs) of launched drugs decreased, reflecting the fact that current drug discovery deviated away from NPs. To change the current situation characterized by high investment but low productivity in drug discovery, efforts should be made to improve the BR of the screening library and hunt drugs more effectively in the biologically relevant chemical space. Additionally, a multiple dimensional molecular descriptor, named the biologically relevant spectrum (BRS) was proposed for quantitative structure—activity relationships (QSAR) study or screening library preparation. Prediction models for 43 biological activity categories were developed with BRS and support vector machine (SVM). In most cases, the overall prediction accuracies were around 95% and the Matthew's correlation coefficients (MCC) were over 0.8. Thirty-seven out of 48 drug-activity associations were successfully predicted for drugs that launched from 2006 to 2012, which were not included in the training data set. A web-server named BioRel (http://ibi.hzau.edu.cn/biorel) was developed to provide services including BR, BRS calculation, activity class, and pharmacokinetic property prediction.

■ INTRODUCTION

Current drug discovery and development are facing an unprecedented predicament where more funds have been invested, but fewer new drugs have been generated.[1] To explore more diverse chemical space, databases of enumerated drug-like chemicals, such as the generated chemical universe database (GDB)[2] and small molecule universe (SMU),[3] have been developed. However, theoretically existing compound space is extremely vast, which is estimated ranging from 10^{23} to 10^{60}.[4−6] Regarding commercially available compounds, there are over 20 million substances in Zinc, which is the largest available chemicals collection.[7] It is impracticable and unnecessary to screen all the available compounds. Therefore, the quality (structural diversity, bioavailability, permeability, drug-likeness, target focusing, etc.) of a screening library is more important than its size.

Although the theoretically existing chemical space is very huge, the chemical space that is actually explored by biological organism is limited. Only a small part of the molecules in GDB or SMU databases are similar to biogenic ligands that can bind to biological macromolecules and then perform a specific function. According to the preferential attachment principle,[8,9] the most basic and ancient biological ligands compose the dominant biogenic chemical space. Therefore, we proposed biological relevance (BR),[10] an index to quantitatively evaluate the possibility of biological origination for a compound. For compounds with the same bioactivity, drugs possess higher average BR than their candidates.[10] The average BR of screening, preclinical compounds, drug candidates in phase I, II, III, and launched drugs increased sequentially from 0.46 to 0.56.[10]

Profiling physicochemical properties and scaffold composition of launched drugs can provide valuable guidance for screening library optimization and improve hit finding efficiency.[11−13] However, comparison of oral drugs launched prior to 1983 and those from 1983 to 2002 showed that the mean value of molecular weight, O and N atoms count, H-bond acceptors, rotatable bonds, and rings count had increased by 13−29%.[14] Recently, Faller et al. compared the property space of new, old drugs, and bioactive molecules and found that traditional drugs occupied only a fraction of the property space of the bioactive molecules.[15] New molecular entities approved (after 2002) are moving away from the chemical space of old drugs (before 2002). The reasons for such property shifting were mainly attributed to the emergence of novel targets,[16−18] the influence of drug-like concepts,[19] nature of high-throughput screening (HTS) hits, and hit-to-lead optimization practices.[20]

Received: July 24, 2013
Published: October 14, 2013

It was suggested that semiempirical rules derived from old drugs were not necessarily valid for new drug discovery.[15] To meet the requirement of new targets, "nondrug-like" chemical space should be explored.[17]

In the past half century, drug discovery strategy was moved from natural products (NPs) inspired physiology screening to combinatorial chemistry (CC) and HTS combination.[21,22] More and more synthetic compounds were used in screening library preparation. Will this strategy change and the accompanying property shifting of the initial screening libraries be the causes of the properties changing of the launched drugs? If so, how should the biologically relevant chemical space be used as guidance for drug discovery? These aroused our interest to address this issue more clearly.

■ MATERIAL AND METHODS

Data Sets for Launched Drugs, Drug Candidates, and NPs. 1614 drugs launched before 2004 were extracted from MDL Drug Data Report (MDDR)[23] with the field "PHASE" annotated with "launched". 242 drugs that approved by U.S. Food and Drug Administration (FDA) in recent years (2005–2012) were extracted from Mullard's series of annual reviews.[24] The structures of these 242 drugs were downloaded from PubChem.[25] Then, the compounds were processed with a Pipeline Pilot[26] protocol. Small fragments were removed. Hydrogen atoms were added to fill the valence. Only organic compounds with appropriate molecular weight (between 80 and 2000 Da) were kept. At last, 1514 drugs were reserved with the time span from 1900 to 2012. The drugs were grouped according to their launching decades. Drugs that launched before 1949 (inclusive) were grouped to "∼1949" and after 2000 (inclusive) were grouped to "2000∼".

Compounds in different development phases (biological testing, preclinical, phase I, phase II, phase III, and launched) were extracted from the MDDR database according to their recordings in "PHASE" field and processed with the same protocol as drug data set preparation.

NPs were extracted from Dictionary of Natural Products (DNP).[27] For each molecule, small fragments and glycosyl groups were removed. Then, inorganic compounds, very small (MW < 80 Da), or very large molecules (MW > 2000 Da) were removed. At last, 189 981 compounds were kept for the following analysis. The similarity matrix between the launched drugs and DNP was calculated with molecular similarity component in Pipeline Pilot. The default algorithm (Tanimoto) and ECFP_4 fingerprint[28] were employed. If the largest similarity between a launched drug and DNP compounds was greater than 0.85, which is a widely accepted threshold,[29,30] the drug was defined as NP-similar.

Physicochemical Properties Calculation. Molecular descriptors (molecular weight, number of atoms, bonds, rings, hydrogen-bond acceptors and donors, etc.) were calculated with Pipeline Pilot. Drug-likeness was calculated with Lipinski's Rule of five (Ro5, molecular weight < 500, logP < 5, the sum of O and N atoms < 10, and the sum of OH and NH groups < 5).[31] All of the statistical analyses were performed with SPSS.[32]

Biological Relevance. First, 2000 molecules were extracted diversely from Kyoto Encyclopedia of Genes and Genomes (KEGG) ligands.[33,34] This set of compounds, named the Biorelevance Representative Compounds Database (BRCD), was used as a standard collection of biologically relevant compounds. During the BRCD preparing process, only the largest fragment of KEGG compound with appropriate molecular weight (between 80 and 2000 Da) was reserved. Inorganic chemicals, polymers, and all known drugs (in-house data set, collected from KEGG drugs, Comprehensive Medicinal Chemistry (CMC),[23] DrugBank,[35] and other available sources) were removed.

Then, based on the similarity array between the objective molecule and BRCD, BR was calculated with the following formula:

$$\mathrm{BR}_3 = S_1 + S_2 + S_3 - S_1S_2 - S_2S_3 - S_1S_3 + S_1S_2S_3 \qquad (1)$$

where, S_i represents the ith largest similarity between the objective molecule and BRCD compounds. By default, BR was calculated at level 3. Similarity calculations for BR scoring were performed with program MOLPRINT 2D using the default settings.[36]

The BR score of a database was calculated as the arithmetic mean of the BR scores of its member compounds. More detailed information on BRCD preparation and BR calculation can be found in our previous paper.[10]

The results reported in this paper are based on an updated version of BRCD (v2011), which was prepared with KEGG ligands that coexisted in the KEGG September 6th, 2006, and March 21st, 2011, releases (v57.0). This update was aimed to remove the compounds that do not belong to the basic metabolic network. Calculation results with the updated BRCD and the original set are comparable, showing the robustness of the algorithm (Supporting Information Figure S1).

Biologically Relevant Spectrum. During the BR calculation process, the similarities between objective compound and BRCD compounds form a 2000-dimensional similarity vector. As BRCD were selected diversely from KEGG ligands, the similarity vector reflected the distribution of the compounds in biological relevant chemical space (sampled by BRCD). Therefore, the vector, named biologically relevant spectrum (BRS), can be used as a multidimensional molecular descriptor in quantitative structure activity relationships (QSAR) study or screening library construction. BRS was calculated with a Perl script, which was modified based on BR calculation script.

Activity Category Similarity Matrix and Prediction Models. Active compounds were extracted from MDDR and classified into 51 categories according the Prous classification system as noted in MDDR "ACTIV_CLASS" field. Two similarity matrixes between these activity categories were calculated. One of them was calculated based on the average BRS of each categories (space center of each activity group) with cosine coefficient. The other one was calculated based on number of shared active compounds with Ochiai coefficient (cosine coefficient for qualitative analysis):

$$\mathrm{corr} = \frac{C}{\sqrt{AB}} \qquad (2)$$

where, C is the number of active compounds shared by two activity categories; A and B are the number of multiactivity compounds in activity categories A and B, respectively.

As a pilot study of BRS' potential application in drug discovery, prediction models for pharmacologically active classes were developed with BRS and support vector machine (SVM). Negative (inactive) compounds were extracted from the Available Chemicals Directory (ACD) database.[23] 2000 MDDR compounds (or 80%, for categories with less than 2000 active compounds) were diversely selected as a positive training set. 5000 diversely selected ACD compounds were treated as a negative training set. Remaining active compounds and 10000

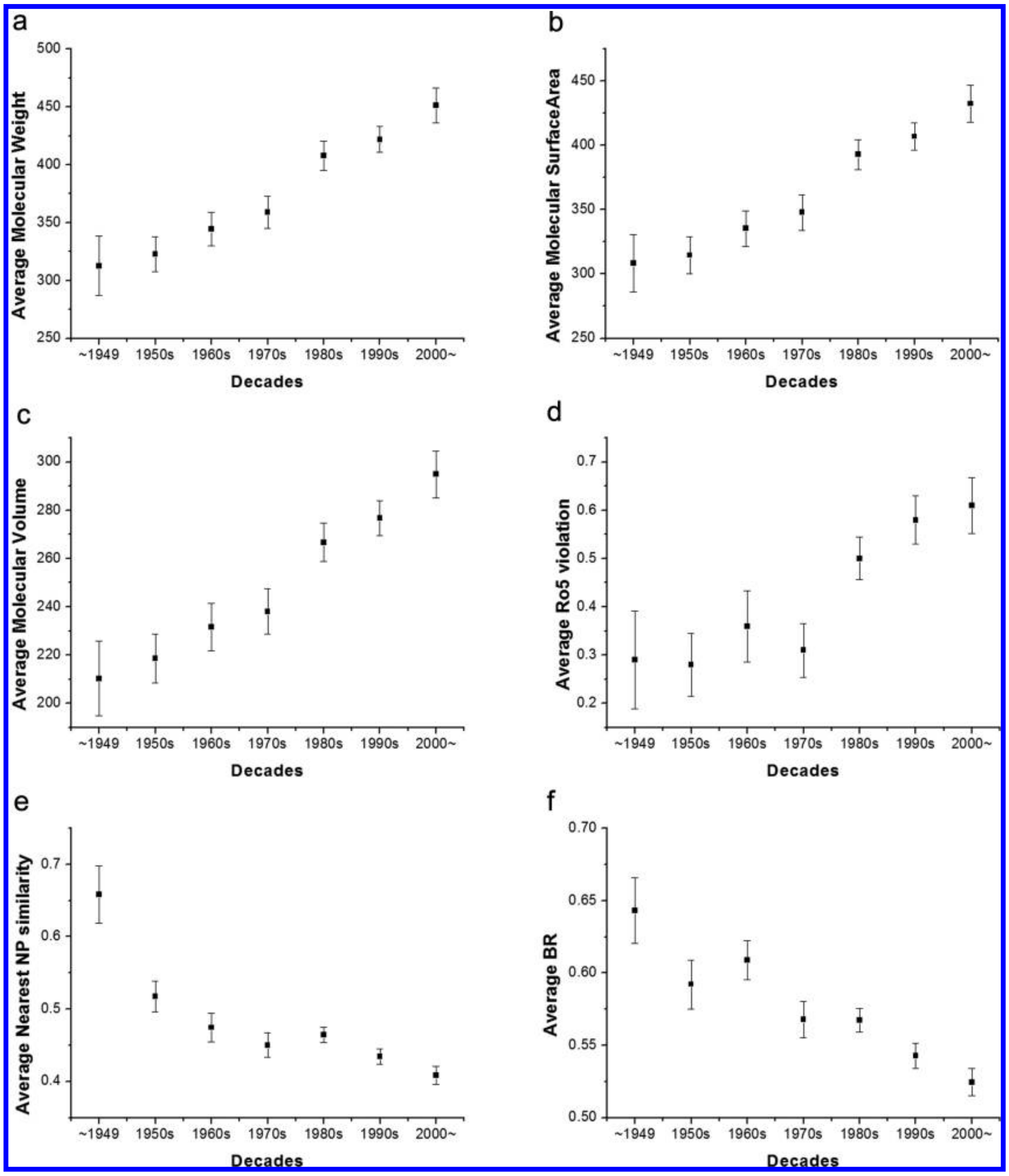


Figure 1. Properties of the launched drugs were shifting decade by decade: (a) average molecular weight; (b) average molecular surface area; (c) average molecular volume; (d) average Lipinski violation number; (e) average nearest NP similarity; (f) average BR score. Standard error was shown with error bars.

randomly selected ACD compounds composed the test set. For classes with less than 1000 compounds in the positive test set, 500 random ACD compounds were used as the negative test set.

SVM and Performance Evaluation. SVM is a supervised learning method for data classification (pattern recognition) and regression introduced by Vapnik and co-workers.[37−39] On the basis of the structural risk minimization (SRM) principle and Vapnik−Chervonenkis (VC) theory, SVM attempts to find an optimal separating hyperplane that provides the minimum number of training errors. Since its excellent performance, SVM has been successfully applied to a wide range of disciplines including quality control and robust regression modeling.

In our study, LIBSVM 3.12[40] with radial basis function (RBF) was used in prediction model developing. Grid search and cross-validation with area under the receiver operating characteristic curve (AUC) were performed to search the optimal parameters. AUC is a widely used benchmarking index for classification model verification. AUC varies from 0.5 (no discriminative power) to 1.0, a higher value indicating a better predictive model. Generally, a model with AUC > 0.9 possesses excellent predictive power.

The following parameters were used to assess the performance of the models: precision, recall, accuracy (ACC), F1 score, Matthew's correlation coefficient (MCC), and AUC.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Table 1. Average BR and Lipinski's Violations of Some Commonly Used Compound Collections

databases[a]	number total	average BR	number BR > 0.5	sum of Ro5 violations	average Ro5 violations	number Ro5 violated
DrugBank	6543	0.585	4255	2638	0.4032	1470
GDB-13	1000000	0.295	33595	115	0.0001	115
CMC	7988	0.547	4893	2517	0.3151	1653
ACD	10000	0.517	5513	3507	0.3507	2254
MDDR	10000	0.488	4074	7300	0.7300	4284
DNP	10000	0.663	8513	5434	0.5434	3391
Launched drugs	1514	0.561	961	710	0.4690	405

[a] Only molecules with number of atoms < 400 and molecular weight < 2000 were analyzed. For GDB-13, ACD, MDDR, and DNP, compounds were selected randomly.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (5)$$

$$\text{F1} = \frac{2\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (7)$$

Here, TP, TN, FP, and FN are the abbreviations for true positive, true negative, false positive, and false negative predictions, respectively. ACC is a measure of true prediction in the whole data set, whereas recall and precision reflect the prediction accuracies for bioactive compounds and predicted active ones, respectively. MCC considers over and under prediction and provides a much more balanced evaluation of prediction. MCC = 1 means a perfect prediction, whereas MCC = 0 indicates a random prediction. The prediction models were optimized and tested using a 5-fold cross validation.

BioRel Server. A web-server named BioRel was designed to facilitate the utilization of BRS and the prediction models. The web interface was written in PHP running on Apache HTTP service. Background calculation was implemented with Perl and Python script. Molecular fingerprint and similarity were calculated with MOLPRINT 2D.[36] Molecular file format conversion and hydrogen adding were performed with Open Babel.[41]

RESULTS AND DISCUSSION

Property Shifting of the Launched Drugs. Over 50 physicochemical properties of the 1514 launched drugs were calculated. Simple statistics were summarized in Supporting Information Table S1 and some were shown in Figure 1. Since the property's distribution patterns are far from normal, both mean and median values of these properties were provided. The mean values were compared with analysis of variance (ANOVA).

As shown in Figure 1 and Supporting Information Table S1, the properties of the launched drugs were chronologically changing decade by decade. The average molecular size (reflected with molecular weight, number of atoms, bonds, surface area, and volume, shown in Figure 1a−c) has increased. Molecular size is an important measure of hit quality. A study by Leeson group showed that the average molecular weight of the drug candidates were decreasing from phase I, II, III clinical test to preregister and marketed drugs.[42] A similar observation was also reported by an earlier study.[43] The increasing molecular size was mainly a result of wide application of HTS and structure-based drug design, which often generate bulky "active" compounds. Besides, current drugs are less drug-like in terms of Ro5, which may be a result of increased molecular weight, HBA, and HBD numbers. Some of the other properties also showed steady trends (maybe with an exception in one or two decades). Most of these results are consistent with earlier reports.[14,15,44]

To illustrate the influence of the initial screening libraries, the similarity matrix between the launched drugs and DNP compounds was calculated. The average nearest NP similarity (the largest similarity value in the similarity vector to NP compounds) dropped from 0.658 (∼1949) to 0.409 (2000∼) with an exception in the 1970s (dropped quickly to 0.450) (Figure 1e). If a drug's nearest similarity to DNP compounds is greater than 0.85, we define this drug as NP-similar.[29,30] The proportion of NP-similar drugs dropped from 32.65% (∼1949) to 6.99% (2000∼), also with an exception in the 1970s (7.89%), as Supporting Information Table S1 shows. The exception in 1970s may be caused by the fact that FDA intensified its efforts to accelerate the approval of more important drugs from the mid-1970s.[45]

BR of the Launched Drugs. The average BR of launched drugs gradually dropped from 0.643 in the first half of the 20th century to 0.524 after 2000 (Figure 1f). The decreasing trend of BR is consistent with NP-similar study and the fact that NPs have gradually been marginalized by major pharmaceutical research companies as sources for new drug discovery in the past 50 years.[21]

The BR score and average Ro5 violations of some commonly used library sources are summarized in Table 1. Although approved drugs hold both high BR score and high drug-likeness, calculations showed no correlation between BR and Ro5 violations.[10] Therefore, a chemical space characterized with BR and drug-likeness (characterized with Ro5 violation) can be constructed to compare the distribution of compounds with different originations. As shown in Figure 2, ACD compounds have the highest drug-likeness and a relatively low BR score, while DNP compounds have the highest biologically relevant but a low drug-likeness. Compounds at biological testing stage in MDDR have the lowest BR and drug-likeness. Compounds at later development phase possess higher BR score and less Ro5 violation. The green arrow shows the direction of current drug optimization, from biological testing, preclinical, phase I, II, and III, to launched drugs. During the development process, both drug-likeness and BR score increase. The average BR score changed sequentially from 0.49 (hits) to

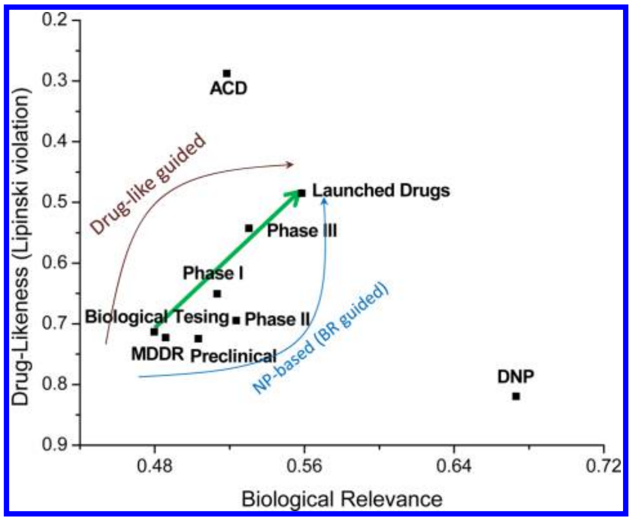


Figure 2. Distribution of compounds of different origination and development phases in the chemical space composed with BR and drug-likeness. DNP, MDDR, and ACD were calculated based on 8000 randomly selected compounds. Glycosyl was not removed for DNP compounds.

0.56 (drugs) (results were calculated with the updated BRCD, also shown in Supporting Information Figure S1b).

Favorable candidates for drug development should possess both high drug-likeness and high BR score (located in the upper-right part in Figure 2). Currently, drug-like was widely accepted and applied in both industrial and academic research. Therefore, efforts should be done to improve the BR score of screening compounds and hunting drugs in biologically relevant chemical space.

Biologically Relevant Chemical Space. Drug discovery is the screening process for druggable compounds in chemical space. The chemical space related to drug discovery (drug-like, biologically relevant, synthetic, and NP) can be illustrated with a Venn diagram in Figure 3. The whole chemical universe includes all the compounds that exist theoretically. Take GDB-13 (a database of theoretically possible compounds with no more than 13 heavy atoms[2]) as an example, almost all compounds in this database are drug-like according to Ro5 (over 99%, calculated with 1 M random subset, Table 1). However, the average BR score of the 1 M GDB-13 subset is only 0.29, much lower than that of CMC and ACD (0.59 and 0.52, respectively). For compounds with only 13 atoms, the average BR score of CMC and ACD databases are 0.55 and 0.53, which are also much higher than GDB-13 compounds. These results suggest that BR can serve as an effective complementary method to drug-likeness in screening library preparation.

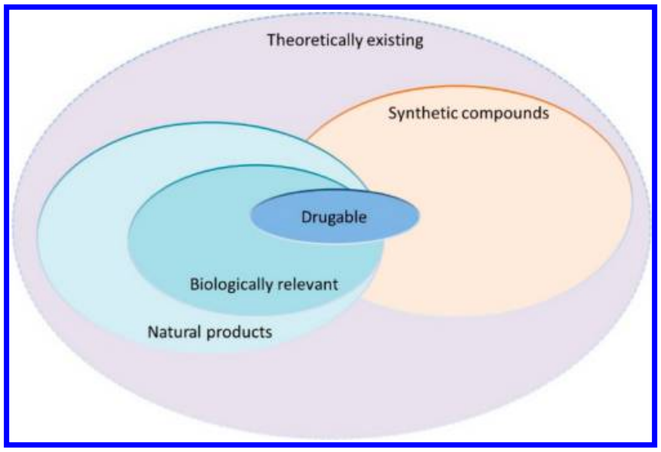


Figure 3. Venn diagram of the distribution of commonly used libraries in chemical space.

For being optimized in a very long natural selection process (biosynthesis, metabolizing, transportation, or utilizing), NPs have the intrinsic superiority to bind with biological macromolecules and thus to be drug candidates. According to earlier statistics, about 50% of modern drugs are NPs or derived from NPs.[46,47] NP-likeness and metabolite-likeness were proposed for virtual screening and library design.[48−50]

Considering the fact that the protein structure is conservative during the long biological evolution process, the number of protein folds is limited. The chemical space exploited by biological organism should also be conservative. This conservative part of NP chemical space, which we named biologically relevant space, includes primary metabolites and the most important secondary metabolites (the core part of NPs). Compared with NP space, biologically relevant space includes only the essential molecules which constitute the basic metabolic requirements of an organism and play irreplaceable roles in biological systems. Analogues of these essential metabolites can inhibit or break the corresponding receptor−ligand reorganization process and be used for chemotherapeutics (e.g., antibiotic or anticancer drug design).[51] More importantly, these metabolites are built with the most basic fragments for life organism systems, which are quite different from organic synthetic compounds.[52−55] Thus, biologically relevant compounds are more convenient to be handled by human metabolic system and have better pharmacokinetic properties than synthetic chemicals. Obviously, drug hunting should be carried out in the biologically relevant space instead of the whole chemical space (GDB or SMU).

Compared with biologically relevant space, the structure and bioactivity diversity of NPs are extremely wide. In addition to essential metabolites, NPs also include other functional secondary metabolites that are generated by organisms for signal transfer, defense against invaders, protecting organisms from oxidation, and living with extreme conditions, etc. In our previous study, the discriminant capability of BR based on NP-BRCD (BRCD constructed with NPs) was worse than BRCD,[10] indicating that although NPs are of great significance for drug discovery, indexes derived from NPs may be not suitable in general drug-likeness rule deducing.

Navigating the Chemical Space with BRS. The similarity vector between the objective compound and BRCD compounds, named BRS, gives the distribution of the objective molecule in biologically relevant chemical space and can be used for chemical space analysis and screening library preparation.

Fifty-one categories of compounds with different activities were extracted from MDDR databases. Then, the BRS of these compounds were calculated. To reduce the descriptor dimension, principal component analysis (PCA) was performed on the chemical space (characterized with BRS vector) center of the 51 activity categories. The first 20 principal components explained over 99% cumulative variances of the total variances (Supporting Information Figure S2). As shown in Supporting

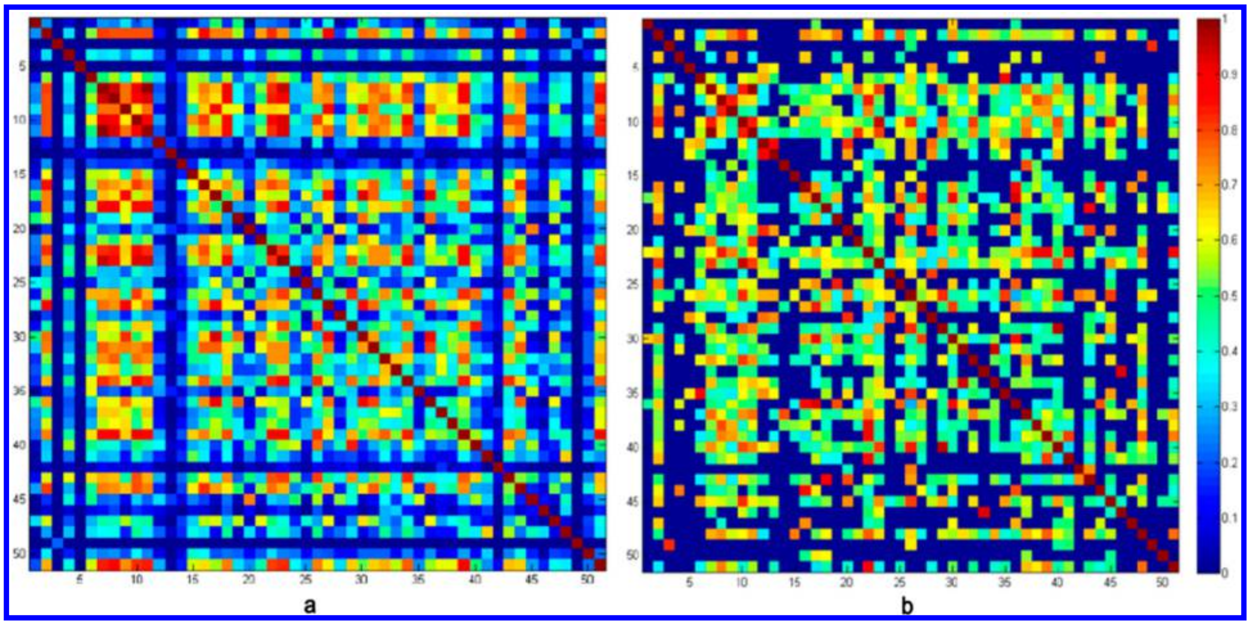


Figure 4. Correlation matrix of the 51 active categories. Similarity was calculated with (a) the average BRS and (b) the number of shared active compounds. Both the x and y axis represent the 51 activity categories. The similarity (cosine coefficient) between the categories were shown with different color from blue to red. Because the shared active compounds between the categories are very limited, to make these heatmaps comparable, the elements of these two matrixes were scaled with (a) a fifteenth power operation and (b) a seventh root operation. The two heatmaps show similar pattern, which implies that BRS can reveal the connections between chemical features and bioactivities.

Table 2. Prediction Results for 40 FDA Approved NMEs

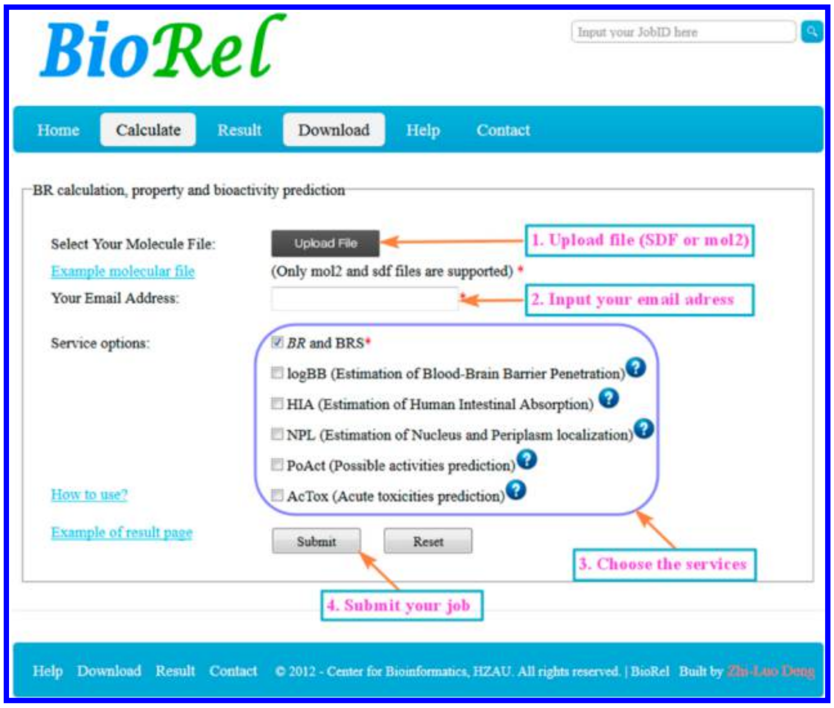
bioactivity[a]	successfully predicted drugs	failed to predict
antineoplastic	Abiraterone; Bosutinib monohydrate; Cabozantinib; Ingenolmebutate; Omacetaxinemepesuccinate; Ponatinib; Regorafenib; Vismodegib; Vemurafenib; Eltrombopagolamine; Dasatinib; Nilotinib hydrochloride monohydrate; Pazopanib hydrochloride	Crizotinib; Enzalutamide
anticoagulant	Apixaban	
antibacterial	Bedaquiline; Dexlansoprazole	
antiviral (AIDS)	Boceprevir; Raltegravir potassium	
anticonvulsant	Gabapentin enacarbil	Perampanel
bronchodilator	Indacaterol; Arformoterol tartrate	
antiparkinsonian	Ioflupane i-123	Perampanel
antidiabetic	Vildagliptin; Saxagliptin	Linagliptin; Sitagliptinphosphatemonohydrate
antipsychotic	Lurasidone; Lisdexamfetaminemesilate; Pazopanib hydrochloride	Armodafinil
estrogen	Oestradiolvalerate	
antiarthritic		Tofacitinib
antidepressant	Dapoxetine; Armodafinil; Lisdexamfetaminemesilate	Varenicline tartrate
platelet	Eltrombopagolamine	
antiulcerative	Revaprazan hydrochloride; Dexlansoprazole	
Anxiolytic		Afobazole
vasodilator		Regadenoson
hyperlipidemic	Choline fenofibrate	
analgesic	Dexlansoprazole	
hypnotic	Armodafinil	

[a]Armodafinil and Dexlansoprazole belong to three activity categories. Pazopanib hydrochloride, Lisdexamfetaminemesilate, Eltrombopagolamine, and Perampanel belong to three categories.

Information Figure S3, the first 20 principal components of the categories were different from each other.

There were 42 477 compounds in MDDR that possess more than one activity categories annotations. This provides us a way to study the relations between the activity categories. Two similarity matrixes between these 51 categories were calculated with the cosine coefficient. One is based on the averaged BRS vector of each activity group (Figure 4a), and the other is based on the number of shared active compounds (Figure 4b). Similar patterns can be observed in these two heatmaps, which indicated that BRS can reveal the implied chemical features of the compounds and can be used to explore chemical space.

For testing the potential application of BRS as a multidimensional descriptor in virtual screening, 51 biological activity class prediction models were developed with MDDR data collection and the SVM method. Statistics of the prediction models are summarized in Supporting Information Tables S2 and S3. These models yield both high precision and



BR calculation, property and bioactivity prediction

Select Your Molecule File: Upload File

Example molecular file

(Only mol2 and sdf files are supported) *

Your Email Address: *

Service options:

☑ *BR* and BRS*

□ logBB (Estimation of Blood-Brain Barrier Penetration)

□ HIA (Estimation of Human Intestinal Absorption)

□ NPL (Estimation of Nucleus and Periplasm localization)

□ PoAct (Possible activities prediction)

□ AcTox (Acute toxicities prediction)

How to use?

Example of result page

Submit Reset

1. Upload file (SDF or mol2)

2. Input your email adress

3. Choose the services

4. Submit your job

Help Download Result Contact © 2012 - Center for Bioinformatics, HZAU. All rights reserved. | BioRel Built by Zhi-Luo Deng

Figure 5. Snapshot of BioRel job submission page. It needs only four steps to submit a job: (1) upload the molecule file (SDF or mol2 format); (2) input the user's e-mail address, which is used to send the results back; (3) choose the models; and (4) click "submit". A job-ID will be provided after submission, which can be used to query the running status and fetch the results when the calculation was finished.

recall ratio. For the test sets, most ACC scores were around 95% and MCC score are over 0.8. Finally, 43 biological activity prediction models (with over 200 positive compounds in training set) were reserved.

To further verify the accuracy and reliability of these prediction models, 40 drugs were tested with the above models. These drugs were approved by FDA in 2006−2012 and were not included in the original MDDR data set. Thirty-seven out of 48 drug−activity associations were predicted correctly (Table 2).

Several absorption, distribution, and acute toxicity prediction models with good prediction accuracy (verified with the test sets), either qualitative discriminations or quantitative regressions, were developed with BRS and SVM. Some of them (listed in Supporting Information Table S4 and Table S5) were integrated into our web-server (discussed below). The others were under optimization and will be provided after accuracy verification. Details of these models will be published elsewhere. These results proved the efficiency of BRS in virtual screening.

Availability and the BioRel Server. BR or BRS calculation is fast and simple. The new BRCD collection in SDF/mol2 format, the Perl calculation script, and all other noncommercial data are freely available from the authors. To facilitate the usage of the algorithms, a user-friendly web-server, named BioRel, was developed. BioRel can be used online at http://ibi.hzau.edu.cn/biorel. Current services include BR and BRS calculation, possible activity prediction (PoAct), and absorption (human intestinal absorption, HIA), distribution (blood−brain barrier penetration, BBB, and nucleus and periplasm localization, NPL), and acute toxicity (AcTox) prediction. The original data sets of these models were collected from several references.[56−58]

As shown in Figure 5, BioRel is easy to use. The input to the BioRel web-server can be a small molecular file in SDF or mol2 format. Hydrogen atoms are not required because they are automatically added before a calculation running. Molecules with explicit lone pair electron are not supported. A submission confirming message, the result files and the URL of the result webpage will be sent to the user by email. The result files include a BRS matrix file and a text file containing BR score and related predictions. These files will be stored on the server for 20 days.

■ CONCLUSION

The wide application of HTS and CC from 30 years ago changed the situation of NP in drug discovery. However, this paradigm of HTS and CC combination did not lead to the anticipated increase in drug productivity.[59] To reduce the high attrition ratio in drug development phases, medicinal chemists proposed several theoretical indexes, such as drug-likeness, lead-likeness, etc. Among them, Ro5, which is the most widely used and the simplest method to predict drug-likeness, has been widely applied. Besides criterions composed with commonly used properties, medicinal chemists also renewed their interests of NP as the source of drug discovery.[60−62] Compared with synthetic compounds, NP occupy a complementary chemical space, which can be characterized with NP-likeness,[48,49] metabolite-likeness,[50,63,64] or a biogenic bias method.[65] Recently, more and more papers referred to biological relevance (or biologically relevant) for screening library optimization.[66−68] In 2009, we proposed a method to calculate biological relevance quantitatively.[10]

In this paper, our calculation showed that the average BR score of the launched drugs decreased decade by decade. The drugs' similarity to NP showed a similar trend. These results confirmed the fact that current drug discovery strategy is deviated away from NP. Poor biological relevance or medicinal relevance[12,67] is one important reason that caused the situation of "more funds, less outcome" in drug discovery. Therefore, to hunt drugs more efficiently, the biologically relevant chemical space should be exploited.

Moreover, to facilitate navigating the chemical space, a multiple molecular descriptor named BRS was proposed for activity class prediction, QSAR study, and screening library preparation. The effectiveness and efficiency of BRS as a multidimensional descriptor for QSAR study and virtual screening were verified with both discriminative and regressive models. These models were available from the BioRel webserver.

BR or BRS are two-dimensional molecular fragment based methods. Thus, activity prediction models developed with BRS are not suitable for scaffold hopping. However, these models can be used for activity focused compound library optimization. On the other hand, the multidimensional descriptor BRS was well-established for multitarget drug design. Currently, a considerable amount of activity data is freely available from National Cancer Institute (NCI), PubChem, or ChEMBL.[69] Many of them are phenotype based screening results. These systems are suitable for analysis with BRS.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1: Properties of the drugs launched by decades. Table S2: Statistics of the active categories prediction results for the training set. Table S3: Statistics of the active categories prediction results for the test set. Table S4: Discriminatory models which were integrated in BioRel. Table S5: Regression models which were integrated in BioRel. Figure S1: Comparison of results based on BRCD 2006 and BRCD 2011 proved the robust of BR algorithm. Figure S2: Explained variances and cumulative explained variances of the first twenty principal components in principle component analysis on the average BRS vector of the 51 active categories. Figure S3: The first twenty principal components of the average BRS vector of the 51 biological active categories. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dxkong@mail.hzau.edu.cn. Tel.: +86-27-8728 0877.

Present Address

$^{\perp}$X.L.: Department of Pharmaceutical Sciences, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from the National Hightech R&D Program of China (2011AA10A204) and the National Natural Science Foundation of China (21075046, 21275061, and 21172124). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors are grateful to Dr. Nicollas Bertero for critically reading the manuscript and making valuable suggestions.

■ REFERENCES

(1) Ruffolo, R. R. Why has R&D productivity declined in the pharmaceutical industry? *Exp. Opin. Drug Discovery* **2006**, *1*, 99−102.

(2) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

(3) Virshup, A. M.; Contreras-Garcia, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.

(4) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(5) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.

(6) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675−679.

(7) Irwin, J. J.; Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(8) Ji, H. F.; Kong, D. X.; Shen, L.; Chen, L. L.; Ma, B. G.; Zhang, H. Y. Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.* **2007**, *8*, R176.

(9) Eisenberg, E.; Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **2003**, *91*, 138701.

(10) Kong, D. X.; Ren, W.; Lu, W.; Zhang, H. Y. Do biologically relevant compounds have more chance to be drugs? *J. Chem. Inf. Model.* **2009**, *49*, 2376−2381.

(11) Ohno, K.; Nagahara, Y.; Tsunoyama, K.; Orita, M. Are there differences between launched drugs, clinical candidates, and commercially available compounds? *J. Chem. Inf. Model.* **2010**, *50*, 815−821.

(12) Lopez-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Expanding the medicinally relevant chemical space with compound libraries. *Drug Discovery Today* **2012**, *17*, 718−726.

(13) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

(14) Leeson, P. D.; Davis, A. M. Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* **2004**, *47*, 6338−6348.

(15) Faller, B.; Ottaviani, G.; Ertl, P.; Berellini, G.; Collis, A. Evolution of the physicochemical properties of marketed drugs: can history foretell the future? *Drug Discovery Today* **2011**, *16*, 976−984.

(16) Dandapani, S.; Marcaurelle, L. A. Grand challenge commentary: Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* **2010**, *6*, 861−863.

(17) Zhao, H. Lead optimization in the nondrug-like space. *Drug Discovery Today* **2011**, *16*, 158−163.

(18) Morphy, R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. *J. Med. Chem.* **2006**, *49*, 2969−2978.

(19) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.

(20) Keseru, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discovery* **2009**, *8*, 203−212.

(21) Carter, G. T. Natural products and Pharma 2011: strategic changes spur new opportunities. *Nat. Prod. Rep.* **2011**, *28*, 1783−1789.

(22) Ortholand, J. Y.; Ganesan, A. Natural products and combinatorial chemistry: back to the future. *Curr. Opin. Chem. Biol.* **2004**, *8*, 271−280.

(23) *MDL databases (CMC, ACD, MDDR, ToxFinder)*, version 2004.1; Elsevier MDL: San Leandro, CA, 2004.

(24) Mullard, A. 2012 FDA drug approvals. *Nat. Rev. Drug Discovery* **2013**, *12*, 87−90.

(25) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* **2008**, *12*, 217−241.

(26) *Pipeline Pilot*, version 8.5; Accelrys: San Diego, CA, 2012.

(27) *Dictionary of Natural Products (DNP)*, version 17.2; Chapman & Hall/CRC Press: London, 2008.

(28) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(29) Chalk, A. J.; Worth, C. L.; Overington, J. P.; Chan, A. W. PDBLIG: classification of small molecular protein binding in the Protein Data Bank. *J. Med. Chem.* **2004**, *47*, 3807−3816.

(30) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(31) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery. Rev.* **1997**, *23*, 3−25.

(32) *SPSS*, version 15.0; SPSS: Chicago, IL, 2006.

(33) Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277−280.

(34) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354−357.

(35) 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.

(36) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(37) Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neura.l Netw.* **1999**, *10*, 988−999.

(38) Sánchez A, V. D. Advanced support vector machines and kernel methods. *Neurocomputing* **2003**, *55*, 5−20.

(39) Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199−222.

(40) Chang, C. C.; Lin, C. J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(41) 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.

(42) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.

(43) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(44) Proudfoot, J. R. The evolution of synthetic oral drug properties. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1087−1090.

(45) Dranove, D.; Meltzer, D. Do Important Drugs Reach the Market Sooner? *Rand. J. Econ.* **1994**, *25*, 402−423.

(46) Newman, D. J. Natural products as leads to potential drugs: an old process or the new hope for drug discovery? *J. Med. Chem.* **2008**, *51*, 2589−2599.

(47) Harvey, A. L. Natural products in drug discovery. *Drug Discovery Today* **2008**, *13*, 894−901.

(48) Jayaseelan, K. V.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinf.* **2012**, *13*, 106.

(49) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.

(50) Dobson, P. D.; Patel, Y.; Kell, D. B. "Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today* **2009**, *14*, 31−40.

(51) Lamichhane, G.; Freundlich, J. S.; Ekins, S.; Wickramaratne, N.; Nolan, S. T.; Bishai, W. R. Essential metabolites of Mycobacterium tuberculosis and their mimics. *MBio* **2011**, *2*, e00301−00310.

(52) Bajorath, J. Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 431−439.

(53) Feher, M.; Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.

(54) Grabowski, K.; Schneider, G. Properties and Architecture of Drugs and Natural Products Revisited. *Curr. Chem. Biol.* **2007**, *1*, 115−127.

(55) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem., Int. Ed.* **1999**, *38*, 643−647.

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

(57) Hou, T.; Wang, J.; Li, 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.

(58) Muehlbacher, M.; Spitzer, G. M.; Liedl, K. R.; Kornhuber, J. Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1095−1106.

(59) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9*, 203−214.

(60) Piggott, A. M.; Karuso, P. Quality, not quantity: the role of natural products and chemical proteomics in modern drug discovery. *Comb. Chem. High Throughput Screening* **2004**, *7*, 607−630.

(61) Paterson, I.; Anderson, E. A. Chemistry. The renaissance of natural products as drug candidates. *Science* **2005**, *310*, 451−453.

(62) Desai, M. C.; Chackalamannil, S. Rediscovering the role of natural products in drug discovery. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 436−437.

(63) Peironcely, J. E.; Reijmers, T.; Coulier, L.; Bender, A.; Hankemeier, T. Understanding and classifying metabolite space and metabolite-likeness. *PLoS One* **2011**, *6*, e28966.

(64) Gupta, S.; Aires-de-Sousa, J. Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. *Mol. Diversity* **2007**, *11*, 23−36.

(65) Hert, J.; Irwin, J. J.; Laggner, C.; Keiser, M. J.; Shoichet, B. K. Quantifying biogenic bias in screening libraries. *Nat. Chem. Biol.* **2009**, *5*, 479−483.

(66) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442−446.

(67) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272−17277.

(68) Eberhardt, L.; Kumar, K.; Waldmann, H. Exploring and exploiting biologically relevant chemical space. *Curr. Drug Targets* **2011**, *12*, 1531−1546.

(69) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−1107.