JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Prediction of Synthetic Accessibility Based on Commercially Available Compound Databases

Yoshifumi Fukunishi,*[,†,‡] Takashi Kurosawa,[‡,§] Yoshiaki Mikami,[‡,§] and Haruki Nakamura[∥]

[†]Molecular Profiling Research Center for Drug Discovery (molprof), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan

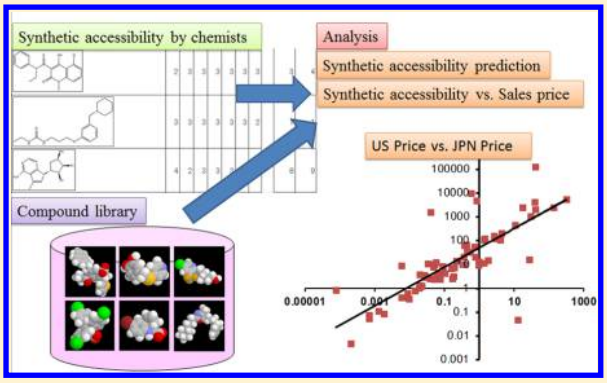
[‡]Technology Research Association for Next-Generation Natural Products Chemistry, 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan

[§]Hitachi Solutions East Japan, 12-1 Ekimae-honcho, Kawasaki-ku, Kawasaki, Kanagawa 210-0007, Japan

[∥]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Supporting Information

ABSTRACT: A compound's synthetic accessibility (SA) is an important aspect of drug design, since in some cases computer-designed compounds cannot be synthesized. There have been several reports on SA prediction, most of which have focused on the difficulties of synthetic reactions based on retro-synthesis analyses, reaction databases and the availability of starting materials. We developed a new method of predicting SA using commercially available compound databases and molecular descriptors. SA was estimated from the probability of existence of substructures consisting of the compound in question, the number of symmetry atoms, the graph complexity, and the number of chiral centers of the compound. The probabilities of the existence of given substructures were estimated based on a compound library. The predicted SA results reproduced the expert manual assessments with a Pearson correlation coefficient of 0.56. Since our method required a compound database and not a reaction database, it should be easy to customize the prediction for compound vendors. The correlation between the sales price of approved drugs and the SA values was also examined and found to be weak. The price most likely depends on the total cost of development and other factors.



1. INTRODUCTION

A compound's synthetic accessibility (SA) is an important aspect of drug design, since in some cases computer-designed compounds cannot be synthesized.[1,2] And while the concept of SA is well-known among medicinal and synthetic chemists, it is less familiar among computational chemists, who generally lack a background in synthetic chemistry. Therefore, although computational chemists often engage in computer-guided drug design, they may find SA estimation of newly designed compounds difficult. An automatic SA evaluation program would be helpful in such cases and should be required when new compounds are generated by a de novo design program. There is thus a great need for an automatic SA prediction program. Accordingly, several computational methods have been developed, such as CAESA,[3] RECAP,[4] WODCA,[5] LHASA,[6] RASA,[7] RSsvm,[8] AIPHOS,[9] and SYLVIA,[10] to perform retrosynthesis[11,12] and/or SA prediction for the compounds in question. In these methods, the synthetic reaction path is predicted by software, and both the difficulty of reactions and the availability of reagents (starting materials) are examined using a database. These methods work well and should be useful in drug design. However, the steric effects (such as atom collisions and other interatomic interactions) were not taken into consideration in these approaches. Also, these SA prediction techniques depend on the availability of reagents, the catalogs of which have changed every year.

The study most closely related to our present study was reported by Takaoka et al.[13] In this study, several chemists evaluated the SA (in the original paper, "synthetic easiness") of each compound in a teaching data set by expert manual assessment. The compounds in the teaching data set were described by molecular descriptors, and the weights of the descriptors were determined to reproduce the SA. This approach worked well, and the report suggested that SA is a useful but not well-defined concept.[14] In the present work, we present a new method for predicting SA based on a commercially available compound database and molecular descriptors. The idea behind our method was very similar to that in a previous study.[15] In both the present and the above-mentioned methods,[13,15] reaction databases and retrosynthesis analyses were not necessary. SA was estimated from the probability of the existence of substructures of the compound calculated based on a compound library, the number of

Received: July 17, 2014
Published: November 24, 2014

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259−3267

symmetry atoms, and the number of chiral centers in the compound. A steric effect, such as that of atomic collision, could be partially taken into consideration by the probability of the existence of substructures of the compound. It has recently become easy to access free compound libraries, such as PUBCHEM,[16] ChEMBL,[17] ZINC,[18,19] and LigandBOX.[20,21] Since our method uses a compound library but not a reaction database, it should be easy for compound vendors to customize their predictions using our approach. Then, the vendors could establish their own libraries with the resulting compounds.

In addition, we examined the relationship between the sales price and SA of approved drugs. Recently, the sales prices of newly approved drugs have increased. The recently developed kinase inhibitors are especially expensive. Of course, the price of a drug is critical to its practical adoption by patients. For this reason, the ability to predict the price of a new drug in the drug-design process would be highly useful. Our hypothesis is that drugs that are difficult to synthesize will be expensive, based on the fact that, in the past, some approved drugs were not actually produced because their sales prices were not high enough to yield profits. We checked this hypothesis in the present study by examining 100 new drugs and 100 generic drugs.

2. METHODS

In the current study, SA was calculated as follows.

$$\mathrm{SA} = c_1 S_{\mathrm{prob}} + c_2 S_{\mathrm{sym}} + c_3 N_{\mathrm{chiral}} + c_4 S_{\text{graph-complexity}} + c_0 \quad (1)$$

$$S_{\mathrm{sym}} = \frac{N_{\mathrm{sym}}}{N_{\mathrm{tot}}} \quad (2)$$

where c_0, c_1, c_2, c_3, and c_4 are fitting parameters. Parameters c_0–c_4 are optimized to reproduce the SA determined by expert manual assessment. In the present study, the SA determined by expert manual assessment is named "human SA" and the SA calculated by eq 1 is named "calculated SA". SA is estimated from the probability of existence (S_{prob}) calculated based on a compound library, the number of symmetry atoms (N_{sym}), the total number of atoms (N_{tot}) of the compound, the number of chiral centers of the compound (N_{chiral}), and the graph complexity ($S_{\text{graph-complexity}}$).[22] N_{sym} is the number of chemically (topologically) equivalent atoms. The importance of symmetry, chirality, and graph complexity was suggested previously.[15]

If a compound consists of fragments frequently found in an available-compound database, it should be easy to synthesize. On the other hand, if a compound consists of fragments rarely or never found in an available-compound database, it would be difficult to synthesize. The probability of existence (S_{prob}) was calculated on the basis of the decomposition of the compound into fragments, and the probability of existence of each fragment was estimated according to the compound library. This idea was introduced previously.[15] We used two fragmentation methods to obtain the molecular descriptors: atom-based fragmentation and bond-based fragmentation (see Figure 1). Our descriptors were very similar to the extended connectivity fingerprint (ECFP) descriptors developed by SciTegic, which are substructural fingerprints.[15]

In atom-based fragmentation, the ith fragment of M order fragmentation consists of the atoms connected to the ith atom by contiguous M bonds.

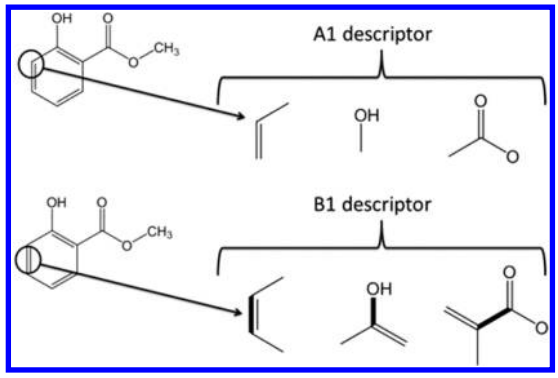


Figure 1. ECFP-like atom-based and bond-based fragmentations.

In bond-based fragmentation, the ith fragment of M order fragmentation consists of the atoms connected to one of the two atoms of the ith bond by contiguous M bonds.

The atom-based and bond-based fragments include the atom-type information and bond-type information, respectively, in addition to the topology, and the definitions of atom and bond types follow those of the Sybyl mol2 file format.

All compounds in the library were decomposed into M order fragments. Let N and $N(i)$ be the total number of N order fragments found in the library and the total number of N order fragments found in the library that were exactly the same as the ith fragment of the compound in question, respectively. The probability of existence of the ith fragment in library ($P(i)$) is given by

$$P(i) = N(i)/N \quad (3)$$

The total probability of existence of the compound in question is then given as

$$P_{\mathrm{total}} = \prod_i P(i) \quad (4)$$

and

$$S_{\mathrm{prob}} = \log_{10}(P_{\mathrm{total}}) = \sum_i \log_{10}(P(i)) \quad (5)$$

The compound library should consist of already-synthesized available compounds.

Chemical reactions occur at chemically equivalent atoms. Thus, it should be easier to synthesize symmetrical substructures than nonsymmetrical substructures. The synthesis of a chiral structure is known to involve a difficult reaction. A previous report showed that the graph complexity was important for evaluating SA. Thus, we examined the N_{sym}, N_{chiral}, and $S_{\text{graph-complexity}}$.

We also calculated the formation heat of the compound in question by using MOPAC PM3 (Quantum Chemistry Program Exchange, (QCPE), Indiana University, Bloomington, Indiana, USA, http://openmopac.net/index.html), since the stability of a compound depends on its formation heat. In general, distorted unstable compounds (such as cubane) should be difficult to synthesize, even if the compound could be synthesized.

In this study, the correlations between two sets of values were calculated (i.e., the correlation between the predicted value and the manual assessment, the correlation between the predicted values, and so on). The all-correlation coefficients (R) were Pearson product-moment correlation coefficients (Pearson correlation coefficients).

3. DATA PREPARATION

Our compound data set consisted of 256 compounds, including 12 from the SYLVIA data,[12] 100 from the RASA data,[9] and 144 randomly selected from the KEGG DRUG database (http://www.genome.jp/kegg/drug/).[23] KEGG DRUG is a free database of approved drugs. For all 256 compounds, 2 chemists from the Nard Institute and 2 chemists from Fujimoto Chemicals evaluated the SA. For the RASA data set, 5 chemists evaluated the SA, and for SYLVIA, 11 chemists evaluated the previously published SA. For the KEGG DRUG data set, the SA was averaged across the results by the four chemists. For the RASA and SYLVIA data sets, we calculated the average SA for all evaluated results. The lists of the human SA values are summarized in the Supporting Information (Table S1).

To calculate S_{prob}, a library of known existing compounds is necessary. We prepared nine libraries to evaluate the library-dependence and size-dependence of our prediction method. Libraries L1, L2, and L3 were nonredundant, and their compounds were randomly selected from the LigandBOX database.[20,21] Each library (L1, L2, and L3) consisted of 100 000 compounds. Libraries M1, M2, and M3 were 50 000-compound libraries randomly selected from L1, L2, and L3, respectively. Libraries S1, S2, and S3 were 10 000-compound libraries randomly selected from M1, M2, and M3, respectively.

The compound data of the libraries (L1−L3, M1−M3, S1−S3) extracted from the LigandBOX database were prepared as three-dimensional (3D) coordinates in the Sybyl mol2 file format; however, only 2D structure data were used in the present study. The atom and bond types also followed the Sybyl mol2 format. The 3D structure of each compound was optimized by energy minimization using the General AMBER force field (GAFF)[24] and myPresto/cosgene[25] with MOPAC AM1 atomic charge. The details of the database were described in our previous papers.[20,21] The 3D structures of the 256-compound set were prepared in the same manner as in the LigandBOX procedure. The 3D structures were prepared to calculate the heat of formation, but in the prediction models (FA1−FA6, FB1, and FB2), only the 2D structures were used.

For the price examination, new drugs and generic drugs were extracted from DrugBank (http://www.drugbank.ca/).[26] About 100 new drugs and 100 generic drugs were arbitrarily selected. When multiple brands were available, the average price was used. The lists of the drugs are summarized in the Supporting Information (Table S2).

4. RESULTS AND DISCUSSION

4.1. Correlation between Human SA and Each Term of the Prediction Model. We prepared a compound database of human SA that consisted of 256 compounds and had 4 chemists from 2 chemical vendors (2 from Nard Institute and 2 from Fujimoto Chemicals) evaluate each compound. The SA of each compound was ranked from 1 (easy to synthesize) to 10 (difficult to synthesize), as in previous works.[9,12,15] The SA of each compound was compared with the calculated compound descriptors. The data sets are described in detail in the Data Preparation section. This data set was used in all validation tests in the present study.

The atom-based S_{prob} values were calculated from 1- to 6-order fragments (A1−A6) and the bond-based S_{prob} values were calculated from 1- to 2-order fragments (B1 and B2). We did not examine further bond-order fragments, since the results were similar to those obtained for the atom-based fragments.

Table 1 shows the correlation between each descriptor and the SA evaluated by the chemists (human SA). The actual values of S_{prob} and the deviations are summarized in Table 2. The three compound libraries (L1, L2, and L3) were used as the database for the atom-based (A1−A6) and bond-based (B1 and B2) descriptors. Most of the descriptors were correlated with the human SA; the exceptions were the atom-based first-order fragment (A1) and the heat of formation. This means that the $P(i)$ value of the ith atom (eq 3) could show the SA around the ith atom. This could tell us which part of the compound in question should be difficult to synthesize. The results obtained by the atomic-based fragmentation were almost the same as those obtained by the bond-based fragmentation. As was the case in the previous reports, the S_{sym}, N_{chiral}, and S_{graph} complexity showed good correlations with the human SA.

Table 1. Pearson Correlation Coefficients between Human SA and Descriptors

	library		
R	L1	L2	L3
S_{sym}	0.267	0.267	0.267
S_{chiral}	0.428	0.428	0.428
heat[a]	0.060	0.060	0.060
S_{graph}	0.345	0.345	0.345
A1[b]	0.118	0.111	0.111
A2[b]	0.332	0.320	0.310
A3[b]	0.500	0.507	0.499
A4[b]	0.494	0.503	0.505
A5[b]	0.446	0.457	0.465
A6[b]	0.404	0.419	0.428
B1[b]	0.402	0.404	0.400
B2[b]	0.474	0.479	0.478

[a]Heat of formation calculated by MOPAC PM3. [b]S_{prob} descriptor.

Table 2. Average S_{prob} Values and Their Standard Deviation over a 256-Compound Set Calculated Using the L1 Library

model	$\log_{10}(S_{prob})$	σ of $\log_{10}(S_{prob})$
A1	−16.7	−60.7
A2	−85.5	−63.3
A3	−134.1	−64.0
A4	−153.7	−64.9
A5	−162.5	−65.0
A6	−168.4	−65.1
B1	−153.3	−71.6
B2	−223.9	−48.8

We examined the correlation between S_{prob} and the human SA using the L1, L2, L3, M1, M2, M3, S1, S2, and S3 libraries, since the atom-based (A1−A6) and bond-based (B1 and B2) descriptors were all correlated highly with human SA. The calculated SA was obtained by using eq 1 with $c_2 = c_3 = c_4 = 0$, and the other two parameters (c_0 and c_1) were optimized. Table 3 shows the computational root-mean-square deviation error (RMSD), the average error, and the correlation coefficient between the human SA values and the SA values calculated by these eight models. The models used were SA1−SA6, SB1, and SB2, and these models employed A1−A6, B1, and B2 as S_{prob}, respectively. The results were obtained by leave-one-out (LOO) cross-validation tests and multiple-linear regression

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259−3267

Table 3. SA Results Calculated by Using Only the S_{prob} Term (Two-Parameter (c_0 and c_1) Model)[a]

library L1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.980	1.555	0.115	1.973	1.549	0.118
A2	4.222	3.910	0.333	4.202	3.893	0.332
A3	1.810	1.411	0.498	1.800	1.405	0.500
A4	1.501	1.165	0.492	1.494	1.160	0.494
A5	1.501	1.177	0.444	1.493	1.171	0.446
A6	1.519	1.200	0.402	1.512	1.194	0.404
B1	1.569	1.167	0.401	1.561	1.162	0.402
B2	1.464	1.108	0.473	1.456	1.103	0.474
library L2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.987	1.564	0.108	1.980	1.558	0.111
A2	4.206	3.843	0.322	4.187	3.828	0.320
A3	1.886	1.493	0.505	1.877	1.486	0.507
A4	1.503	1.151	0.501	1.495	1.146	0.503
A5	1.491	1.153	0.455	1.483	1.148	0.457
A6	1.500	1.160	0.417	1.493	1.155	0.419
B1	1.589	1.181	0.403	1.581	1.175	0.404
B2	1.466	1.107	0.478	1.458	1.102	0.479
library L3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.991	1.570	0.108	1.984	1.564	0.111
A2	4.188	3.787	0.311	4.169	3.772	0.310
A3	1.938	1.532	0.497	1.928	1.526	0.499
A4	1.517	1.160	0.503	1.509	1.154	0.505
A5	1.489	1.148	0.463	1.482	1.143	0.465
A6	1.494	1.154	0.425	1.486	1.149	0.428
B1	1.607	1.191	0.399	1.598	1.185	0.400
B2	1.469	1.105	0.476	1.461	1.100	0.478
library M1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.086	1.641	0.093	2.079	1.635	0.095
A2	3.746	3.303	0.364	3.736	3.293	0.363
A3	1.609	1.237	0.487	1.600	1.232	0.489
A4	1.498	1.170	0.454	1.491	1.164	0.456
A5	1.511	1.186	0.412	1.504	1.181	0.415
A6	1.538	1.211	0.374	1.531	1.205	0.376
B1	1.573	1.168	0.390	1.565	1.163	0.392
B2	1.469	1.116	0.459	1.462	1.110	0.461
library M2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.080	1.639	0.083	2.073	1.633	0.086
A2	4.012	3.672	0.355	3.998	3.659	0.352
A3	1.647	1.284	0.511	1.638	1.278	0.513
A4	1.461	1.128	0.490	1.454	1.123	0.492
A5	1.483	1.151	0.439	1.476	1.145	0.441
library M2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A6	1.509	1.171	0.399	1.502	1.166	0.401
B1	1.585	1.177	0.402	1.577	1.171	0.403
B2	1.463	1.108	0.472	1.455	1.103	0.473
library M3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.080	1.640	0.081	2.073	1.634	0.084
A2	4.084	3.781	0.352	4.068	3.767	0.350
A3	1.686	1.311	0.518	1.677	1.305	0.519
A4	1.454	1.116	0.506	1.447	1.111	0.508
A5	1.474	1.134	0.453	1.467	1.129	0.456
A6	1.496	1.153	0.411	1.489	1.148	0.414
B1	1.601	1.185	0.400	1.592	1.179	0.401
B2	1.462	1.100	0.477	1.454	1.095	0.478
library S1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.729	2.129	0.014	2.721	2.122	0.013
A2	1.908	1.507	0.448	1.899	1.500	0.450
A3	1.478	1.146	0.453	1.470	1.140	0.455
A4	1.510	1.185	0.397	1.503	1.180	0.399
A5	1.540	1.209	0.360	1.533	1.203	0.362
A6	1.559	1.224	0.336	1.552	1.219	0.339
B1	1.559	1.158	0.390	1.552	1.153	0.392
B2	1.469	1.129	0.444	1.462	1.124	0.446
library S2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.539	1.990	0.016	2.532	1.983	0.013
A2	2.258	1.828	0.439	2.248	1.821	0.438
A3	1.460	1.116	0.493	1.452	1.111	0.495
A4	1.477	1.145	0.434	1.470	1.140	0.436
A5	1.511	1.179	0.391	1.505	1.174	0.393
A6	1.534	1.193	0.361	1.527	1.188	0.363
B1	1.568	1.161	0.407	1.560	1.155	0.408
B2	1.461	1.108	0.463	1.454	1.103	0.464
library S3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	2.544	1.986	0.022	2.537	1.980	0.019
A2	2.317	1.885	0.446	2.307	1.878	0.445
A3	1.466	1.120	0.497	1.458	1.115	0.499
A4	1.465	1.133	0.446	1.458	1.128	0.448
A5	1.509	1.173	0.393	1.502	1.168	0.395
A6	1.535	1.196	0.361	1.528	1.191	0.363
B1	1.564	1.155	0.414	1.555	1.150	0.415
B2	1.454	1.101	0.471	1.447	1.095	0.472

[a]MLR: multilinear regression. R: Pearson correlation coefficient.

(MLR) over the data of all 256 compounds using the L1, L2, L3, M1, M2, M3, S1, S2, and S3 libraries. In the leave-one-out cross-validation test, 1 datum was selected as the test datum to be predicted, and the other data were used as teaching data to generate the prediction model equation. The test data were selected one after another in a given data set until all data had been selected as test data. The results obtained by the atom-based fourth-order fragmentation (SA4) were better than those

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259−3267

Table 4. SA Results Calculated by Equation 1 (Five-Parameter (c_0–c_4) Model)[a]

library L1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.357	1.049	0.516	1.326	1.026	0.516
A2	1.338	1.023	0.534	1.308	1.002	0.535
A3	1.316	1.011	0.555	1.287	0.990	0.556
A4	1.309	1.001	0.561	1.280	0.980	0.562
A5	1.324	1.018	0.546	1.296	0.997	0.547
A6	1.336	1.034	0.536	1.307	1.013	0.536
B1	1.353	1.043	0.519	1.323	1.020	0.520
B2	1.331	1.015	0.540	1.302	0.994	0.541
library L2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.357	1.050	0.516	1.326	1.027	0.516
A2	1.338	1.027	0.534	1.309	1.006	0.534
A3	1.312	1.008	0.558	1.283	0.987	0.559
A4	1.311	0.997	0.560	1.281	0.977	0.561
A5	1.323	1.013	0.548	1.294	0.992	0.549
A6	1.333	1.028	0.539	1.303	1.007	0.540
B1	1.354	1.044	0.519	1.323	1.022	0.519
B2	1.332	1.019	0.539	1.303	0.998	0.540
library L3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.358	1.050	0.515	1.327	1.028	0.515
A2	1.339	1.027	0.533	1.310	1.005	0.533
A3	1.312	1.007	0.558	1.283	0.986	0.559
A4	1.307	0.996	0.563	1.278	0.975	0.564
A5	1.319	1.005	0.552	1.290	0.985	0.553
A6	1.329	1.019	0.542	1.300	0.997	0.543
B1	1.354	1.045	0.518	1.324	1.023	0.519
B2	1.333	1.020	0.539	1.304	0.998	0.539
library M1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.358	1.050	0.515	1.327	1.027	0.515
A2	1.342	1.026	0.530	1.312	1.005	0.531
A3	1.325	1.013	0.546	1.296	0.992	0.547
A4	1.327	1.012	0.543	1.298	0.991	0.545
A5	1.336	1.026	0.534	1.307	1.005	0.536
A6	1.347	1.044	0.524	1.317	1.022	0.525
B1	1.355	1.044	0.517	1.324	1.021	0.518
B2	1.337	1.019	0.535	1.307	0.998	0.536
library M2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.357	1.049	0.516	1.326	1.026	0.516
A2	1.339	1.028	0.533	1.310	1.007	0.533
A3	1.313	1.011	0.558	1.284	0.990	0.559
A4	1.310	0.999	0.561	1.281	0.978	0.562
A5	1.324	1.015	0.547	1.295	0.995	0.548
A6	1.336	1.034	0.535	1.307	1.013	0.536
B1	1.353	1.044	0.519	1.323	1.021	0.519
B2	1.333	1.021	0.539	1.304	1.000	0.539
library M3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.357	1.050	0.515	1.327	1.027	0.516
A2	1.338	1.025	0.535	1.308	1.004	0.535
A3	1.306	1.003	0.564	1.277	0.982	0.565
A4	1.299	0.988	0.571	1.270	0.968	0.572
A5	1.317	1.005	0.554	1.288	0.984	0.555
A6	1.331	1.025	0.541	1.301	1.003	0.542
B1	1.354	1.045	0.518	1.324	1.023	0.519
B2	1.331	1.018	0.540	1.302	0.997	0.541
library S1	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.358	1.048	0.515	1.326	1.025	0.516
A2	1.338	1.022	0.533	1.309	1.001	0.534
A3	1.324	1.016	0.547	1.294	0.995	0.549
A4	1.336	1.033	0.536	1.305	1.011	0.538
A5	1.351	1.050	0.522	1.320	1.028	0.522
A6	1.359	1.059	0.514	1.327	1.036	0.515
B1	1.355	1.044	0.517	1.325	1.021	0.518
B2	1.334	1.015	0.537	1.304	0.994	0.539
library S2	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.356	1.048	0.516	1.326	1.026	0.517
A2	1.339	1.027	0.533	1.309	1.005	0.533
A3	1.317	1.015	0.554	1.288	0.994	0.555
A4	1.326	1.017	0.545	1.296	0.995	0.547
A5	1.342	1.039	0.530	1.312	1.017	0.531
A6	1.352	1.051	0.521	1.321	1.029	0.522
B1	1.354	1.044	0.519	1.323	1.022	0.519
B2	1.334	1.021	0.538	1.305	1.000	0.538
library S3	cross-validation test			MLR		
level	RMSD	average error	R	RMSD	average error	R
A1	1.356	1.048	0.517	1.325	1.026	0.517
A2	1.333	1.021	0.539	1.304	0.999	0.539
A3	1.310	1.007	0.562	1.280	0.985	0.563
A4	1.312	1.001	0.559	1.282	0.980	0.561
A5	1.336	1.029	0.536	1.306	1.008	0.537
A6	1.349	1.048	0.524	1.318	1.026	0.525
B1	1.353	1.044	0.520	1.322	1.021	0.520
B2	1.327	1.014	0.545	1.298	0.993	0.545

[a]MLR: multilinear regression. R: Pearson correlation coefficient.

obtained by the other fragmentations and were similar to those obtained by the bond-based second-order fragmentation. The correlation coefficients between the predicted values and the human SAs were about 0.5 for models SA4 and SB2.

4.2. SA Prediction by the Theoretical Model and Its Accuracy. We examined the prediction accuracy of eq 1 with all c_0–c_4 parameters free and with the use of L1, L2, L3, M1, M2, M3, S1, S2, and S3 libraries. Eight prediction models

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259–3267

(FA1−FA6, FB1, and FB2) were examined. In each model, S_{prob} was calculated by each of the A1−A6, B1, and B2 descriptors. Table 4 shows the computational root-mean-square deviation error (RMSD), the average error, and the correlation coefficient between the human SAs and the values calculated by these eight models (FA1−FA6, FB1, and FB2). The results were obtained by LOO cross-validation tests and the MLR across the data of all 256 compounds using the L1, L2, L3, M1, M2, M3, S1, S2, and S3 libraries, just as in the above examinations of models SA1−SA6, SB1, and SB2.

The results obtained by the atom-based fourth-order fragmentation were better than those obtained with the other fragmentations and similar to those obtained by the bond-based second-order fragmentation. The correlation coefficients between the predicted SA values and the human SAs were about 0.56 for each of models FA3 and FA4. The computational results did not depend on the size of the libraries used for the calculation of the descriptors, and a library containing 10 000−100 000 compounds was large enough for prediction. Since our method required a compound library but not a reaction database, it should be easy to customize predictions for compound vendors by using this method. Compound vendors could make their own libraries of the compounds they have made. Figure 2 shows the correlations between the predicted values obtained by model FA4 with library L1 and the synthetic accessibilities evaluated by the chemists. The correlation graphs obtained using libraries L2 and L3 were similar to those in Figure 1, and they were omitted. There were clear correlations between the calculated and the human SAs.

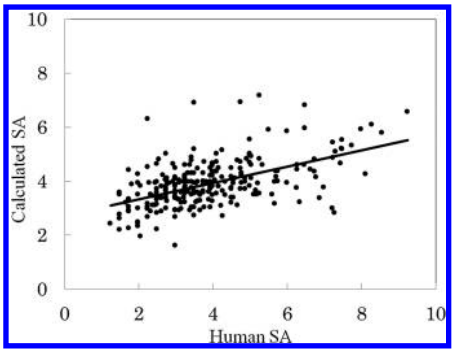


Figure 2. Correlation between the human SA and calculated SA obtained by the cross-validation test by model FA4 based on library L1. The solid line shows a fitted curve.

Table 5 shows the correlations between the SAs calculated by models SA1−SA6, B1, and B2 (only S_{prob} was used for the prediction). The SA values were calculated for the 256-compounds set as well as for the above procedures. Since the larger fragments include the smaller fragments, the predicted values were correlated with each other. In particular, the SA4 and FB2 models showed that all correlation coefficients were larger than 0.35, while some models did not correlate at all—for example, there were no correlations between SA1 and SA3, between SA1 and SB1, or between SA2 and SA6. Tables 3 and 4 show that model FA4 was suitable for SA prediction, because model FA4 was able to yield results similar to those of the other models and maintained a high prediction accuracy.

Table 5. Pearson Correlation Coefficient between SAs Calculated Using Only the S_{prob} Term (Two-Parameter (c_0 and c_1) model)

	A1	A2	A3	A4	A5	A6	B1	B2
A1	1.00	0.58	0.04	0.38	0.54	0.66	0.07	0.35
A2		1.00	0.75	0.47	0.30	0.15	0.61	0.52
A3			1.00	0.91	0.79	0.69	0.78	0.86
A4				1.00	0.97	0.91	0.80	0.94
A5					1.00	0.98	0.77	0.92
A6						1.00	0.72	0.89
B1							1.00	0.91
B2								1.00

In Figure 2, the gradient of the fitted line was 0.32 and this gradient was lower than the ideal gradient, 1. The fitted gradients did not change depending on the choice of the compound libraries (S1, S2, S3, M1, M2, M3, L1, L2, and L3). The reason for the low gradient could be the lack of descriptors to evaluate the SA of the test set compounds. In other words, the diversities of the compound libraries were relatively low comparing to the test set. This idea is supported by the weak dependence of the calculated SA on the library size in Table 4. One of the reasons of the weak dependence of result (R and error) on the library size could be low diversity of the compound libraries, and the $P(i)$ values in eq 3 did not change by the choice of the library. The prediction could be improved by using a compound library of rich of diversity.

We also applied a 2-fold cross-validation test. In this test, the 256-compounds set was arbitrarily dived into two groups. One group was used as the teaching set and the other was used as the test set. The test was repeated by exchanging the teaching and test sets, and an average of the two predictions was used as the result. The 2-fold cross-validation tests were performed three times, using model FA4 with the L1, L2, and L3 libraries, respectively. For these three tests, the correlation coefficient between the predicted SA values and the human SA values was 0.40, and the average error of each cross-validation test was 1.20. When the number of compounds in the teaching set was reduced by half, the current prediction method still worked, although the prediction accuracy decreased.

The LigandBOX database (the current version as of June 2013) consists of 4 586 865 compounds supplied by 44 compound vendors. The biggest three libraries are Enamine (1 883 713 compounds), Vitas (1 163 246 compounds), and TimTec (1 066 286 compounds). The average number of entries per library across all vendors is 179 748 compounds. However, if the biggest three vendors are excluded, the average number of entries per library drops to 92 576 compounds. Indeed, 31 out of 44 vendors (70%) supply fewer than 92 000 compounds. Because our prediction method can work with a compound database of smaller than 100 000 compounds, these vendors could use our method to conduct an easy and automatic SA evaluation based on their own compound libraries before their manual assessments.

4.3. Correlation among Human SAs. Table 6 shows the correlations among the human SAs. As shown in previous reports, the human SAs were strongly dependent on the skill and experience of the individual chemist. Thus, the simple average of the correlation coefficients (R) was 0.586. The results obtained by the two chemists who belonged to the same company were similar to each other. The results obtained by the chemists from different companies were different from each other. SA would thus be expected to depend on the equipment available at the individual company, as well as on the training and experience of the company chemists. This would explain

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259−3267

Table 6. Pearson Correlation Coefficient between Human SAs Evaluated by Two Chemists[a]

	FMP1	FMP2	NARD1	NARD2	RASA	SYLVIA
FMP1	1.000	0.931	0.572	0.557	0.355	0.369
FMP2	0.931	1.000	0.552	0.530	0.285	0.413
NARD1	0.572	0.552	1.000	0.862	0.474	0.944
NARD2	0.557	0.530	0.862	1.000	0.438	0.923
RASA	0.355	0.285	0.474	0.438	1.000	N.D.
SYLVIA	0.369	0.413	0.944	0.923	N.D.	1.000

[a]FMP: Fujimoto Chemicals. NARD: Nard Institute.

why the R of our human SA values was much lower than those reported in the previous studies (R = 0.7−0.8).[9,12] If, in fact, our method was not effective at SA prediction, then the correlation coefficient between the calculated SA and the human SA would be expected to be much lower than that between human SAs. However, the R value (0.56) and the average error (1.2) obtained by the FA4 model were similar to those between the human SAs (R = 0.59 with a standard deviation of 0.22 and average error = 1.1). Thus, our SA prediction method was considered to function effectively.

4.4. Relationship between the Sales Price of Approved Drugs and SA. Figure 3 shows the correlation between the logarithm of the sales price of a newly approved drug in the United States (US) and the calculated SA value obtained by model FA4. The drug price is the price per mole. The price showed a weak correlation to the calculated SA, with a correlation coefficient R of 0.32. The correlation coefficients between the sales prices of approved drugs/generics and the SA values are summarized in Table 7. The prices per weight, mole, and tablet did not correlate with the SA values. The logarithms of the prices showed a weak correlation to the calculated SA values. The prices of new drugs in the United States and Japan showed a weak correlation to the calculated SA values. Their correlation coefficients were 0.32 and 0.29, respectively. The prices of generic drugs showed the same trends as the new drugs. The reason for this should be that the price of a generic drug follows the price of the original drug. The average SA of a new drug was 5.6, slightly higher than that of a generic drug (average SA = 5.5). We also examined the list of the correlation coefficients between the sales prices and the SA values (Table 8), broken down by drug. The R values of antiacute and anticancer drugs in Japan showed weak positive correlations, while the others did not show correlation. In the United States, some R values became negative.

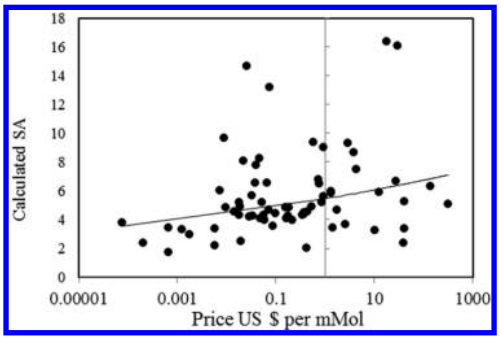


Figure 3. Correlation between the sales price of approved drugs in the US and SAs. The sales price of an approved drug is the price per mole of a new drug in the US. The solid line shows a fitted curve.

Table 7. Pearson Correlation Coefficient between the Sales Prices of Approved Drugs/Generics and the SA Values

US	new approved drug	generic drug
price per weight (mg)	0.03	0.11
price per mole	0.02	0.11
Log(price per mol)	0.32	0.37
Japan	new approved drug	generic drug
price per weight (mg)	0.05	0.27
price per mole	0.12	0.43
Log(price per mole)	0.29	0.27

Table 8. Pearson Correlation Coefficient (R) between the Sales Prices of Newly Approved Drugs and Their SA Values

		US		Japan	
category		number of drugs	R	number of drugs	R
dosage form	topical drug (/mg)	9	−0.45	14	−0.02
	topical drug (/mmol)	9	−0.38	14	0.00
	injection drug (/mg)	13	−0.09	20	0.03
	injection drug (/mmol)	13	−0.08	20	0.19
	oral drug (/mg)	48	0.08	66	0.03
	oral drug (/mmol)	48	0.14	66	0.24
pathophysiology	antichronic disease drug (/mg)	73	0.08	52	0.02
	antichronic disease drug (/mmol)	73	0.15	52	0.04
	antiacute disease drug (/mg)	7	−0.02	7	0.38
	antiacute disease drug (/mmol)	7	−0.02	7	0.43
	anticancer drug (/mg)	11	−0.07	11	0.31
	anticancer drug (/mmol)	11	−0.02	11	0.46

The approved drug sets included some natural compounds, such as steroids, oligosaccharides, macrolides, and other large cyclic compounds. These natural compounds are extremely difficult to synthesize, but they are still relatively inexpensive. However, the current teaching sets for human SA evaluations consist of only non-natural compounds. The difference in the data sets could be one of the reasons why the correlation is poor.

The correlation coefficients between the sales prices of drugs in the United States and Japan are summarized in Table 9. The

Table 9. Pearson Correlation Coefficient (R) between Sales Prices of Drugs in the United States and Japan

	newly approved drugs	generic drugs
price/mg	0.14	0.81
price/mmol	0.14	0.84
Log(price/mg)[a]	0.77	0.63
Log(price/mmol)[a]	0.79	0.66

[a]Correlation coefficient between the logarithms of the drug prices in the United States and the logarithms of the prices in Japan.

dx.doi.org/10.1021/ci500568d | J. Chem. Inf. Model. 2014, 54, 3259−3267

correlation between the logarithm of the price of drugs in the US and the price in Japan was very high (R = about 0.8), even though drug prices in the United States and Japan are controlled by their respective governments, as shown in Table 8.

The price of a drug may depend on many kinds of costs, including the costs of development, phase trials, and patents. This means that a drug that is difficult to synthesize is not always expensive. The average SA of a drug is about 6, meaning that the average difficulty of drug synthesis is not extremely high.

4.5. Miscellaneous Discussion on the Prediction Model. The predicted SA strongly depended on S_{prob}. If the compound library consisted of approved drugs, S_{prob} reflected the specific features of those drugs. For example, the probability of finding a toxic substructure (like −COH, −N−N−, or −O−O−) should be low. Thus, S_{prob} should be a measure of drug-likeness. Since drug-likeness is not a well-defined idea, we did not examine in detail the behavior of S_{prob} values in the present study.

The calculated SA results obtained by model FA4 were not substantially better than those obtained by model SA4. This means that the $S_{graph\text{-}complexity}$, S_{chi}, and S_{sym} terms do not improve the results dramatically. The substructure information includes some information about chirality and symmetry. For example, the symmetry of $-CH_2$ or $-CH_3$ is represented by the A1−A6, B1, and B2 descriptors. If four atoms connected to a center C atom are all different from each other, the center C atom must be a chiral center. This information is also included in the A1−A6, B1, and B2 descriptors. In practical use, the calculated SA results obtained by model FA4 were less dependent on the size of the compound library compared to the calculated SA results obtained by model SA4. Thus, model FA4 should be more useful than model SA4.

The prediction accuracy of our model (R = 0.56 by model FA4) was not superior to the prediction results obtained by the previous studies. One of the possible explanations for these homogeneous results is that the concept of SA remains unclear, and thus there is a limit to the prediction accuracy. Users will tend to choose the computer software most suitable for their purpose among the various types of SA prediction methods.

Our method expends about 1 min in generating a histogram of the substructures present in a library of 10 000 compounds (S1, S2, and S3), and the calculation of eq 1 for each compound costs about 0.01 s on a Xeon 5400 processor (3 GHz). The computational time required for histogram generation of the substructures of a library was almost proportional to the number of compounds in the library.

5. CONCLUSION

We proposed a new method for predicting SA. The input compound was decomposed into ECFP-like substructures, and the probabilities of the substructures were estimated using a database search on a compound library of 10 000−100 000 available compounds. In addition, the symmetry, the number of chiral centers, and the graph complexity of the compounds were considered to predict the SA. The predicted SA showed good agreement with the SA estimated by chemists, with a correlation coefficient of 0.56. Considering that the correlation coefficient between SAs estimated by the chemists was around 0.58, the computational prediction worked well. Model FA4 showed the best prediction accuracy among the eight tested models, and the most important descriptor was the probability of finding the substructure in a given compound library. The number of compounds in a teaching library could be small; namely, 10 000−100 000 compounds were sufficient to calculate the SA. Thus, it should be easy to customize predictions for compound vendors using this method. These vendors could establish their own compound libraries made up of the compounds they have produced. The source code of the program will be freely available from our website (http://presto.protein.osaka-u.ac.jp/myPresto4/index.php?lang=en).

In our approach, the major contribution to the calculation of the SA value was made by S_{prob}, which was calculated by using a group contribution method. This means that the probability of the existence of a fragment structure around each atom could reveal the partial SA of that structure. Thus, it would be possible to determine which part of a given compound would be difficult to synthesize.

We also examined the correlation between drug prices and SA. For both new and generic drugs, the price did not depend on SA and thus presumably depended on the total development cost and other factors. Therefore, the theoretical prediction of drug prices is likely to be very difficult.

■ ASSOCIATED CONTENT

Supporting Information

Table S1: the human SA values with the compound structures. Table S2: the sales prices of approved drugs with the compound sctructures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +81-3-3599-8290. Fax: +81-3-3599-8099. E-mail: y-fukunishi@aist.go.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from the National Institute of Advanced Industrial Science and Technology (AIST), the New Energy and Industrial Technology Development Organization of Japan (NEDO), and the Ministry of Economy, Trade, and Industry (METI), of Japan. Y.F. thanks Dr. Isao Yasumatsu for helpful suggestions.

■ REFERENCES

(1) Schneider, G.; Fechner, U. Computer-based de novo Design of Drug like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(2) Baber, J. C.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681−692.

(3) Gillet, V.; Myatt, G.; Zsoldos, Z.; Johnson, A. SPROUT, HIPPO and CAESA: Tool for de novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discovery Des.* **1995**, *3*, 34−50.

(4) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. RECAP-retrosynthetic Combinatorial Analysis Procedure: a Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.

(5) Pförtner, M.; Sitzmann, M. *Handbook of Cheminformatics*; Gasteiger, J., Ed., Wiley-VCH: Weinheim, 2003; pp 1457−1507.

(6) Johnson, A.; Marshall, C.; Judson, P. Starting Material Oriented Retrosynthetic Analysis in the LHASA Program. 1. General description. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 411−417.

(7) Huang, Q.; Li, L.-L.; Yang, S.-Y. RASA: a Rapid Retrosynthesis-based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. *J. Chem. Inf. Model.* **2011**, *51*, 2768−2777.
(8) Podolyan, Y.; Walters, M. A.; Karypis, G. Assessing Synthetic Accessibility of Chemical Compounds Using Machine Learning Methods. *J. Chem. Inf. Model.* **2010**, *50*, 979−991.
(9) Funatsu, K.; Sasaki, S. Computer-assisted Organic Synthesis Design and Reaction Prediction System, "AIPHOS. *Tetrahedron Comput. Methodol.* **1988**, *1*, 27−38.
(10) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311−325.
(11) Corey, E. J.; Wipke, W. T. Computer-assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178−192.
(12) Timothy, D.; Salatin, T. D.; Jorgensen, W. L. Computer-assisted Mechanistic Evaluation of Organic Reactions. 1. overview. *J. Org. Chem.* **1980**, *45*, 2043−2057.
(13) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-likeness and Ease of Synthesis Using a Data Set in which Compounds are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.
(14) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47*, 4891−4896.
(15) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.
(16) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Bryant, S. H. PubChem: a Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(17) Gaulton, A.; Bellis, L. J.; Bentro, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, P. ChEMBL: a Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.
(18) Irwin, J. I.; Shoichet, B. K. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Info. Model.* **2005**, *45*, 177−182.
(19) Irwin, J. I.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC − a Free Tool to Discover Chemistry for Biology. *J. Chem. Info. Model.* **2012**, *52*, 1757−1768.
(20) Fukunishi, Y.; Sugihara, Y.; Mikami, Y.; Sakai, K.; Kusudo, H.; Nakamura, H. Advanced in-silico Drug Screening to Achieve High Hit Ratio - Development of 3D-compound Database. *Synthesiology* **2008**, *2*, 64−72.
(21) Kawabata, T.; Sugihara, Y.; Fukunishi, Y.; Nakamura, H. LigandBox - a Database for 3D Structures of Chemical Compounds. *BIOPHYSICS.* **2013**, *9*, 113−121.
(22) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Sci.* **1981**, *103*, 3599−3601.
(23) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for Integration and Interpretation of Large-scale Molecular Data Sets. *Nucleic Acids Res.* **2012**, *34*, D109−D114.
(24) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(25) Fukunishi, Y.; Mikami, Y.; Nakamura, H. The Filling Potential Method: a Method for Estimating the Free Energy Surface for Protein-ligand Docking. *J. Phys. Chem. B* **2003**, *107*, 13201−13210.
(26) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−D1097.