http://www.jstage.jst.go.jp/browse/biophysics

BIOPHYSICS

*Databases and Computer Programs*

**Vol. 9, pp. 113–121 (2013)**
**doi: 10.2142/biophysics.9.113**

# LigandBox: A database for 3D structures of chemical compounds

Takeshi Kawabata[1], Yusuke Sugihara[2], Yoshifumi Fukunishi[3] and Haruki Nakamura[1]

*[1]Institute for Protein Research, Osaka University, 3-2, Yamadaoka, Suita, Osaka, 565-0871, Japan*
*[2]Fujitsu Kyushu R&D Center, Life Science Systems Dept., Fujitsu Kyushu Systems Limited, 2-1, Momochihama 2-Chome, Sawara-ku, Fukuoka, 814-8589, Japan*
*[3]Molecular Profiling Research Center for Drug Discovery (molprof), AIST, 2-3-26, Aomi, Koto-ku, Tokyo, 135-0064, Japan*

Received February 20, 2013; accepted July 11, 2013

**A database for the 3D structures of available compounds is essential for the virtual screening by molecular docking. We have developed the LigandBox database (http://ligandbox.protein.osaka-u.ac.jp/ligandbox/) containing four million available compounds, collected from the catalogues of 37 commercial suppliers, and approved drugs and biochemical compounds taken from KEGG_DRUG, KEGG_COMPOUND and PDB databases. Each chemical compound in the database has several 3D conformers with hydrogen atoms and atomic charges, which are ready to be docked into receptors using docking programs. The 3D conformations were generated using our molecular simulation program package, *myPresto*. Various physical properties, such as aqueous solubility (LogS) and carcinogenicity have also been calculated to characterize the ADME-Tox properties of the compounds. The Web database provides two services for compound searches: a property/chemical ID search and a chemical structure search. The chemical structure search is performed by a descriptor search and a maximum common substructure (MCS) search combination, using our program *kcombu*. By specifying a query chemical structure, users can find similar compounds among the millions of compounds in the database within a few minutes. Our database is expected to assist a wide range of researchers, in the fields of medical science, chemical biology, and biochemistry, who are seeking to discover active chemical compounds by the virtual screening.**

Corresponding author: Takeshi Kawabata, Institute for Protein Research, Osaka University, 3-2, Yamadaoka, Suita, Osaka, 565-0871, Japan.
e-mail: kawabata@protein.osaka-u.ac.jp

**Key words:** drug, virtual screening, molecular docking, maximum common substructure

Databases for chemical compounds are essential for drug discovery and chemical biology; however, until a few years ago, most of the chemical databases were not free, and charged fees for accesses. The lack of a free online database has been a barrier to drug discovery by academic researchers[1,2]. However, in 2004, the PubChem database was launched as an open repository for chemical structures and their biological test results[3]. Shortly thereafter, other free databases, such as ChEMBL[4] and ChemSpider (http://www.chemspider.com), were released. Meanwhile, virtual screening approaches are being widely used to discover new biologically active compounds from a database of compounds using various computational techniques[5–8]. They have been classified into two classes; ligand-based and receptor-based screenings. The ligand-based screening uses a 2D or 3D chemical structure of known active compound to retrieve other potential active compounds from a database using similarity measures. The receptor-based screening uses a 3D structure of receptor protein, into which compounds from the database are docked and ranked using potential energy functions. The 3D ligand-based screening and receptor-based screening require 3D conformations of library compounds; however, most of the free chemical databases mainly collect 2D structures of chemical compounds.

To address this issue, we developed the new Web database LigandBox (LIGANDs DataBase Open and eXtensible), containing purchasable chemical compounds for the purpose of drug development, particularly *in-silico* drug docking studies[9]. The compounds are mainly from the suppliers' catalogues, and each compound is characterized by its 3D conformers, electrical charges, and calculated physical properties. They are ready to be used for docking calculation with the standard docking programs, such as *UCSF DOCK*[10], *AutoDock*[11], and *sievgene*[12]. We released the first version of the LigandBox database in 2004, and distributed it by sending DVDs to researchers who requested them[9]. In April 2012, we released the LigandBox Web server, after obtaining the suppliers' permissions to open their compound data.

Several 3D chemical compound databases were also developed. At this moment, the largest and the most popular 3D database is the ZINC database[13,14]. It contains over twenty million commercially available compounds with their 3D conformations and electrical charges, for docking calculation using the UCSF DOCK program[10]. Other 3D databases, such as MMsINC[15] and CoCoCo[16] were also developed. Recently, the PubChem database generated 3D conformers of 92.3% of all the compounds[17]. The basic concept of LigandBox is similar to those of the other 3D compound databases, but our Web database has four characteristics. First, the 3D conformations and the associated property data in LigandBox are produced and maintained by our program package *myPresto*, whose sources are freely downloadable from our WEB site. In contrast, most of the other databases employ commercial programs, such as *CORINA*[18] and *OMEGA*[19]. Second, our database provides several unique properties for the selection, such as aqueous solubility (LogS) and carcinogenicities, which are not described in other databases. These values are useful to select potential active compounds from a huge compound library before starting docking calculation. Third, our LigandBox server provides a chemical search engine for the maximum common substructure (MCS) using our *kcombu* program[20]. Most chemical databases provide a descriptor and substructure searches; however, MCS searches are rarely implemented in them. The similarities detected by the MCS search are more intuitively understood because one-to-one atom correspondences of two chemical structures are explicitly shown. This search engine is useful to extract new potential active compounds structurally similar to known active compounds. The program *kcombu* is also freely downloadable from our Web site. Fourth, LigandBox contains about one million unique compounds that are not registered in the ZINC and PubChem databases, although our database is smaller than them.

## Material and Methods

### Sources of chemical compounds

The chemical compound data were obtained from the catalogue of Namiki Shoji Co., Ltd., the KEGG DRUG / KEGG COMPOUND databases[21], and the ligands of PDB database. The Namiki Shoji catalogue contains more than four million compounds from 37 suppliers. The catalogue is provided twice a year and the LigandBox data are updated once a year. When the LigandBox data are updated, all of the compounds in the previous version are discarded; those in the latest catalogue are newly registered. Some of the compounds in LigandBox may become stock-out between the updating times. The KEGG DRUG database stores the approved drugs in Japan, USA and Europe. The KEGG COMPOUND database stores a collection of small biochemical molecules relevant to biological systems. These data are provided in an SDF or mol file format, with the 2D coordinates (X and Y coordinates), but often lack hydrogen atoms. The ligands of PDB are downloaded in an SDF file format with ideal 3D coordinates from the RCSB PDB Web site. The numbers of 2D molecules and 3D molecules are summarized in Table 1. The source suppliers and databases of the LigandBox compounds are summarized in Table 2.

**Table 1** Number of 2D and 3D structures in the LigandBox database

| | |
|---|---:|
| Number of 2D chemical structures | 4,196,995 |
| Number of 3D chemical structures | 7,025,536 |
| Number of Suppliers and Databases | 40 |

**Table 2** Source suppliers and databases of the LigandBox database

| Name | #2D | Name | #2D | Name | #2D | Suppliers | #2D |
|---|---:|---|---:|---|---:|---|---:|
| ENAMINE | 1790960 | Labotest | 106039 | TOSLab | 17572 | RareChemicals | 9535 |
| Vitas_M | 1056604 | Maybridge | 55804 | Peakdale | 14643 | KEGG_DRUG* | 7283 |
| UOS | 680332 | Synthon_Labs | 49726 | INNOVAPHARM | 14310 | VillaPharma | 6728 |
| TimTec | 451543 | ScientificExchange | 47537 | PDB* | 13974 | WuxiAppTec(Natural) | 5498 |
| Princeton | 417465 | Pharmeks(Natural) | 40742 | Princeton(Natural) | 13258 | Menai | 4872 |
| Asinex | 383666 | Bionet | 39169 | CHEM-X-INFINITY | 13148 | AnalytiConMEGx | 4711 |
| Pharmeks | 367858 | MDD | 31136 | MDPI | 13036 | ChemOvation | 2125 |
| LifeChemicals | 343831 | WuxiAppTec | 30210 | Florida | 12236 | InFarmatik | 1462 |
| OTAVA | 168008 | Intermed | 29296 | KEGG_COMPOUND* | 11780 | Bahrain | 949 |
| CBI | 125592 | AnalytiConNATx | 23278 | Vitas_M(Natural) | 9722 | ChemOvation(Natural) | 55 |

#2D: Number of 2D compounds from each supplier or database.
*: Names of databases. Compounds from these three databases may not be commercially available. Names without asterisks are suppliers.

### Procedures to generate 3D conformations

We will briefly explain the six step procedures to generate the 3D conformations. Further details are available in our previous report[9]. **(i) Removing counter ions:** The counter ions of the chemical compounds are removed. If the molecule consists of two or more connected components, then only the largest connected component is kept. **(ii) Adding hydrogen atoms:** After removing the counter ions of the chemical compounds, the *Hgene* program in *myPresto* is used to predict the missing hydrogen atoms from the bond connection table. The dominant ion form at pH 7 of each compound is prepared (for example $R\text{-}COO^-$, $R\text{-}NH_3^+$) for the protein-compound docking study, while the undissociated form is provided in many compound database (for example R-COOH, $R\text{-}NH_2$). **(iii) Assignment of force field parameters:** the *tplgeneL* program in *myPresto* assigns the energy parameters of the molecular force field. We employ the general AMBER force field (GAFF)[22]. **(iv) Energy optimization to generate 3D conformations:** the *cosgene* program in *myPresto* is utilized for energy optimization. **(v) Modification of chiralities and generation of isomers:** if the chiralities of the generated conformation are not consistent with those specified in the original SDF file, then the conformation is modified. **(vi) Computation of atomic charges by quantum mechanical calculation:** the atomic charges are the Mulliken population obtained by the MOPAC AM1 model (http://openmopac.net/index.html). The source codes of our molecular simulation program package, *myPresto*, are downloadable from the Web site (http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html). It takes about one minute to generate a 3D structure, about a few minutes to calculate physical properties for one chemical compound, using the standard Xeon CPU with one core. A few months are required to calculate 3D structures and physical properties for the four million compounds using about 100 CPU cores.

### Calculation of Physical Properties

Several properties are calculated to annotate the physical and the ADME-Tox (adsorption, distribution, metabolism, excretion and toxicity) properties of each molecule. The properties stored in LigandBox are summarized in Table 3. The molecular weight, molecular charge, number of hydrogen-bond donors, number of hydrogen-bond acceptors and number of chiral atoms are calculated using our in-house programs. The energies of LUMO and HOMO are calculated using the MOPAC AM1 model, and these values are closely related to the photosensitivity. The LogS value of the aqueous solubility is one of the most important values for the ADME property. Compounds with poor solubility tend to have poor absorption, low stability, fast clearance and non-specific binding[23]. The LogS value is often regarded as a similar measure to the LogP value, which is more widely used than the LogS value. Other free compound databases, such as ZINC and PubChem, store the LogP values, because the LogP value is more easily observed by experiment and more easily predicted than the LogS value. However, the LogP is physically different from the LogS value. The LogP value is defined as the partitioning coefficient between *n*-octanol and water, has been accepted as the measure of the

**Table 3** Physical properties stored in the LigandBox database

| Name of properties | Descriptions | Min | Mean | Max |
|---|---|---|---|---|
| MOLECULAR WEIGHT | Molecular mass weight (Da). | 17.0 | 380.3 | 2222.6 |
| MOLECULAR CHARGE | Total charge of molecule. | −12 | 0.12 | 12 |
| NUMBER OF DONOR | Number of hydrogen bond donor atoms. Oxygen or nitrogen bonded with hydrogen atom. | 0 | 1.27 | 45 |
| NUMBER OF ACCEPTOR | Number of hydrogen bond accepor atoms. Oxygen or nitrogen with lone pairs and fluorine. | 0 | 3.90 | 49 |
| NUMBER OF CHIRAL ATOMS | Number of chiral atoms, such as asymmetric carbon atoms. | 0 | 0.43 | 51 |
| LUMO | Energy of the lowest unoccupied molecular orbital (eV), calculated by MOPAC AM1. This value is closely related to photosensitiviy. | −41.5 | −1.3 | 35.2 |
| HOMO | Energy of the highest occupied molecular orbital (eV), calculated by MOPAC AM1. This value is closely related to photosensitivity. | −51.8 | −9.3 | 23.7 |
| LOGS | Value of the aqueous solubility predicted by the machine learning method using features of chemical groups and the MD simulation[24]. A compound with a higher LOGS is more soluble in water. | −30.6 | −4.5 | 376.4 |
| AGGREGATOR PROBABILITY | Estimated based on the LOGS value[24]. A compound with a higher aggregation probability tend to more aggregate. It also reflects non-specific binding ability of the compound. | 0.00 | 0.53 | 1.00 |
| CARCINOGENICITY_FN_AD | A measure of applicability domain for CARCINOGENICITY_FN, calculated by ADMEWORKS. | 0.0 | 454.0 | $4.2\times10^8$ |
| CARCINOGENICITY_FP_AD | A measure of applicability domain for CARCINOGENICITY_FP, calculated by ADMEWORKS. | 0.0 | 454.0 | $4.2\times10^8$ |
| CARCINOGENICITY_FN | Carcinogenicity predicted to reduce false negatives, using ADMEWORKS. | −: 32.5%, +: 62.5% | | |
| CARCINOGENICITY_FP | Carcinogenicity predicted to reduce false positives, using ADMEWORKS. | −: 98.2%, +: 1.8% | | |

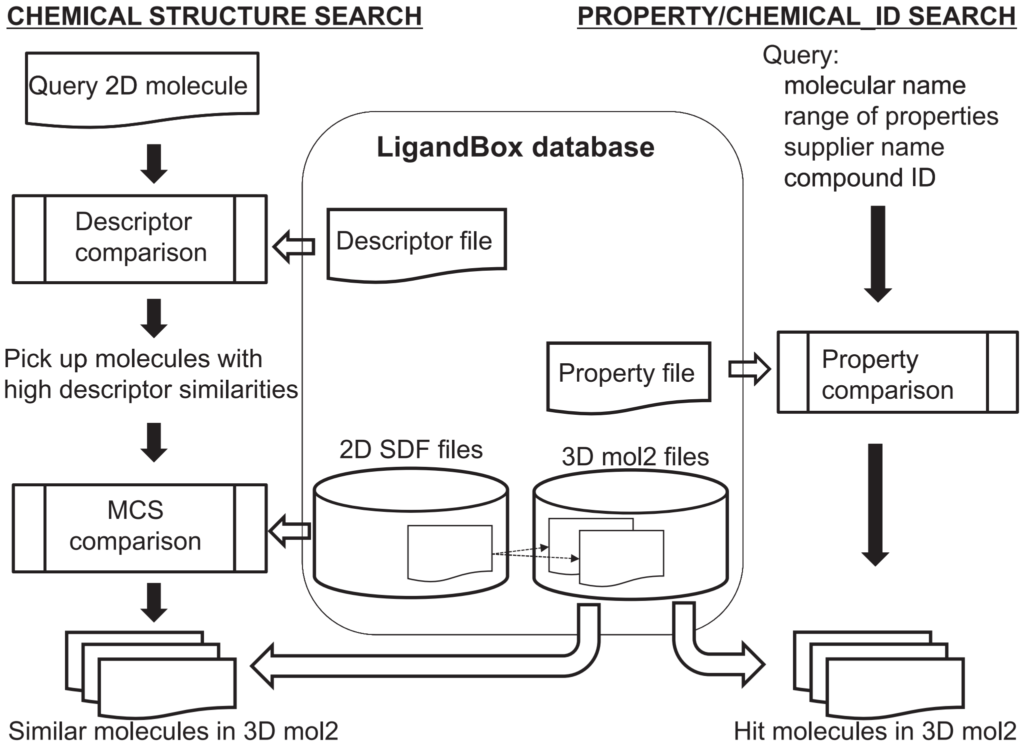


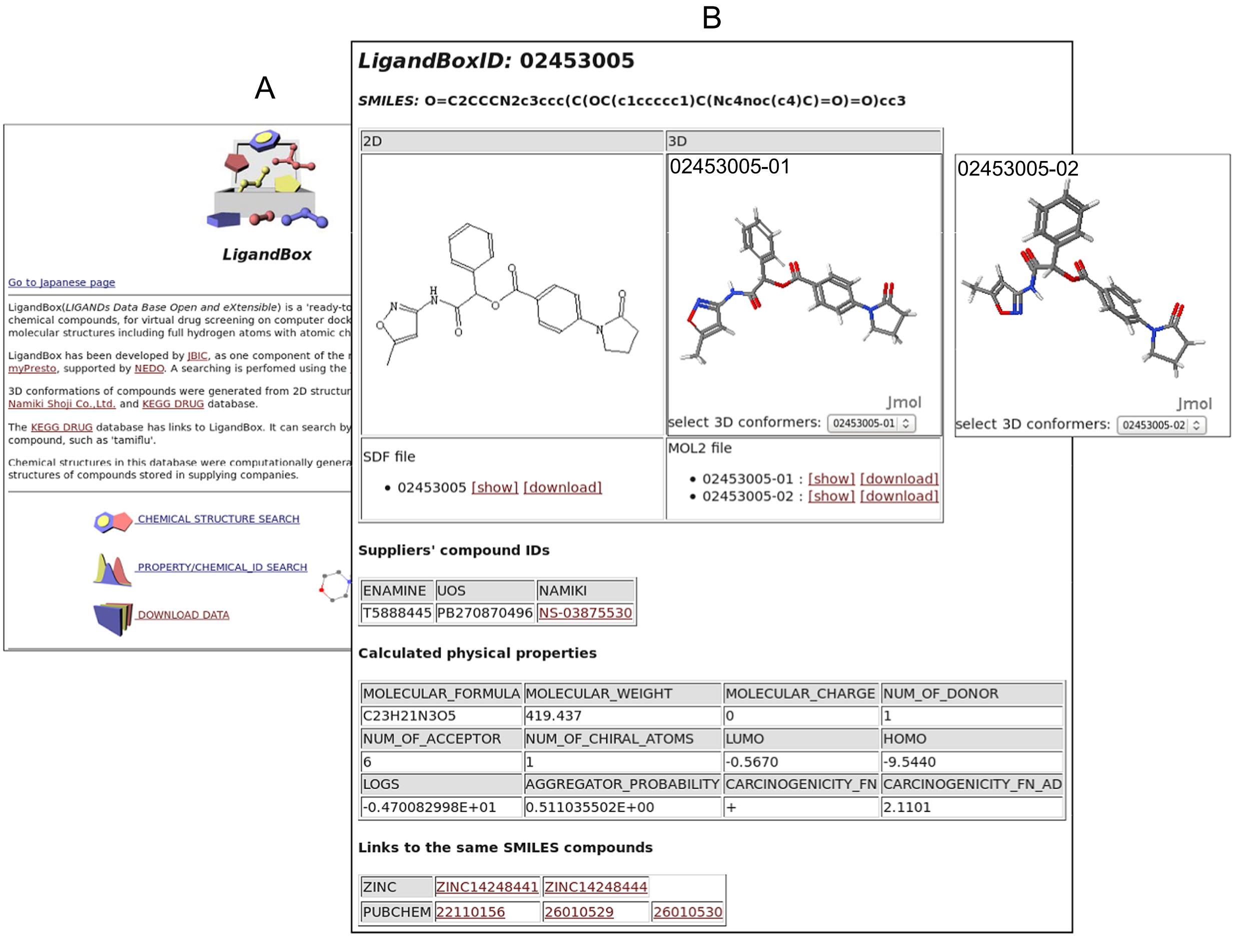
**Figure 1** Schematic view of the LigandBox database.

lipophilicity or membrane permeability. In other words, LogS includes the free energy for melting a crystal of compound, whereas LogP does not. Our LogS value is predicted by the machine learning method based on chemical substructure descriptors and properties derived from MD simulations[24]. The aggregation probability is estimated based on the LogS value, and this probability reflects the non-specific binding of the compound. The carcinogenicity values are predicted using Fujitsu's ADMEWORKS program ver. 6.0 (http://jp.fujitsu.com/group/kyushu/en/services/admeworks/), which was developed based on ADAPT system (ADAPT Computational Chemistry Software Rev. 3.0, Molecular Design Ltd, San Leandro, CA, USA). ADMEWORKS predicts the properties by a multi-linear regression method, based on values of topological, geometric, physicochemical and substructure descriptors derived from the molecular structures.

### Outline of the database system

Figure 1 shows the outline of our database system. The top Web page is shown in Figure 2A, and an example of a molecule stored in the database is shown in Figure 2B. Each chemical compound has its own LigandBox ID number (such as 02453005) and its own single 2D SDF file. Its corresponding multiple conformers are provided as 3D mol2 files, with an extended ID number, such as 02453005-01 and 02453005-02. The 2D and 3D structures and physical properties are shown on one page. If the compounds have corresponding entries in the PubChem[3], ChEMBL[4] and ZINC[14] databases, then the links to these databases are also shown. The KEGG_DRUG[21] and PDBj databases[25] have made links to LigandBox for the corresponding entries on their WEB pages. The unique SMILES string is also shown, which is calculated using Weiningers' algorithm[26,27] implemented in our *kcombu* program[20]. By copying and pasting a SMILES string, a 2D molecular structure in LigandBox can be easily exported to other programs.

As shown in Figure 1, we prepare two files for the searching: the property file and descriptor file. The property file contains all of the calculated physical properties, molecular names, and chemical IDs for all of the 2D molecules. The descriptor file is for the chemical structures search, containing the atom-pair descriptors for all of the molecules. For the property/chemical ID search, a simple in-house program searches the property file by queries of molecular names, chemical IDs, and upper and lower values of physical properties and the hit molecules are shown. These hit molecules can be downloaded in a 3D MOL2 file format. A random selection of hit molecules is also available. This function is useful to generate a decoy dataset for testing performances of virtual screening methods. For the structure search, a descriptor comparison for the descriptor file is performed before the MCS comparison, as described in the next section. The machine of the WEB server has eight CPU cores. One job for the chemical structure search or property/chemical ID search uses only one CPU core.



Go to Japanese page

LigandBox(*LIGANDs Data Base Open and eXtensible*) is a 'ready-to
chemical compounds, for virtual drug screening on computer dock
molecular structures including full hydrogen atoms with atomic ch

LigandBox has been developed by JBIC, as one component of the
myPresto, supported by NEDO. A searching is perfomed using the

3D conformations of compounds were generated from 2D structur
Namiki Shoji Co.,Ltd. and KEGG DRUG database.

The KEGG DRUG database has links to LigandBox. It can search by
compound, such as 'tamiflu'.

Chemical structures in this database were computationally genera
structures of compounds stored in supplying companies.

CHEMICAL STRUCTURE SEARCH

PROPERTY/CHEMICAL_ID SEARCH

DOWNLOAD DATA

**LigandBoxID: 02453005**

**SMILES: O=C2CCCN2c3ccc(C(OC(c1ccccc1)C(Nc4noc(c4)C)=O)=O)cc3**

select 3D conformers: 02453005-01

select 3D conformers: 02453005-02

SDF file

- 02453005 [show] [download]

MOL2 file

- 02453005-01 : [show] [download]
- 02453005-02 : [show] [download]

**Suppliers' compound IDs**

| ENAMINE | UOS | NAMIKI |
|---|---|---|
| T5888445 | PB270870496 | NS-03875530 |

**Calculated physical properties**

| | | | |
|---|---|---|---|
| MOLECULAR_FORMULA | MOLECULAR_WEIGHT | MOLECULAR_CHARGE | NUM_OF_DONOR |
| C23H21N3O5 | 419.437 | 0 | 1 |
| NUM_OF_ACCEPTOR | NUM_OF_CHIRAL_ATOMS | LUMO | HOMO |
| 6 | 1 | -0.5670 | -9.5440 |
| LOGS | AGGREGATOR_PROBABILITY | CARCINOGENICITY_FN | CARCINOGENICITY_FN_AD |
| -0.470082998E+01 | 0.511035502E+00 | + | 2.1101 |

**Links to the same SMILES compounds**

| | | | |
|---|---|---|---|
| ZINC | ZINC14248441 | ZINC14248444 | |
| PUBCHEM | 22110156 | 26010529 | 26010530 |

**Figure 2** A. top page of the LigandBox database. B. An example page for a molecule. We show the molecule with the LigandBox ID = 02453005, provided by two suppliers, ENAMINE and UOS. This molecule has one chiral carbon, but its chirality is not clearly described in its SDF file provided in the supplier's catalogue. Two 3D conformations, 02453005-01 and 02453005-02, are generated in the 3D mol2 format. 3D conformations are visualized by Jmol (http://www.jmol.org).

### Chemical Structure Search

The chemical structure search is performed by the combination of a descriptor search and the 2D maximum common substructure (MCS) search. The MCS is defined as a maximum substructure present in two molecules with the same atom types and bond connections. Similarities detected by the MCS are more intuitively understood than the descriptor search, because corresponding atom pairs between two molecules are explicitly indicated as shown in Figure 3. However, because the MCS requires a large computational cost, we employ the descriptor search as a preliminary filter. As the first step of the search, the descriptor search using the atom-pair descriptors quickly finds similar compounds to the query molecule structures. The atom-pair descriptors encode all pairs of heavy atoms in a molecule together with the atom types and the length of the topological distance (shortest path distance) between them[28]. An example of the atom pair descriptor is shown in Figure 4. Second, the MCS comparisons are performed against the 2D SDF file library, for the restricted number of similar compounds found by the descriptor search. The maximum number of compounds for the MCS calculation is restricted to 1,000 as the default. It is because the MCS comparison is much slower than the descriptor search, although the MCS provides a one-to-one atom correspondence between two chemical compounds. The *kcombu* program is used for the MCS calculation, based on the build-up heuristic algorithm[20]. Various types of common substructure searches are available through the Web server, including isomorphic, substructure, connected MCS (C-MCS), and topologically-constrained disconnected MCS (TD-MCS). TD-MCS is a disconnected MCS allowing only the $\theta$ difference in the topological distance of the corresponding atom pairs ($\theta$=0,1,2). It can sensitively detect weak similarities, such as compound pairs with different-length linker atoms. Examples of C-MCS and TD-MCS results are shown in Figure 3.

The Web server accepts a query chemical structure in various ways: uploading SDF/MOL2/PDB files from the user's local computer, pasting a SMILES string and drawing a molecular structure by the JME editor[29], as shown in Figure

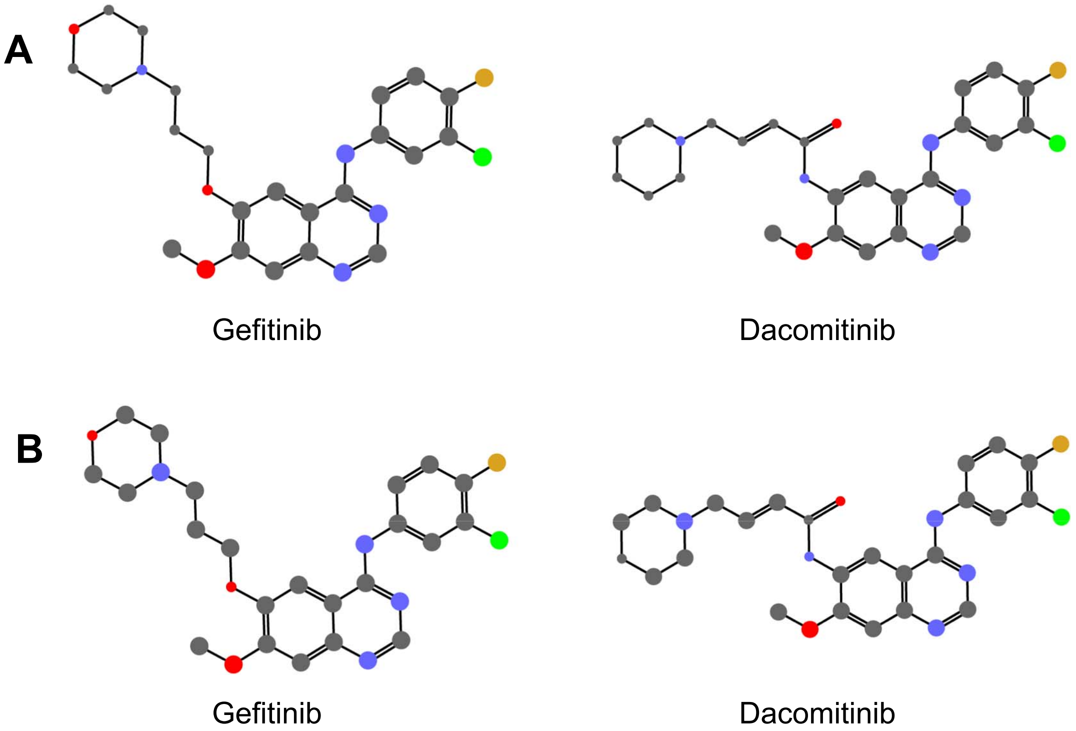
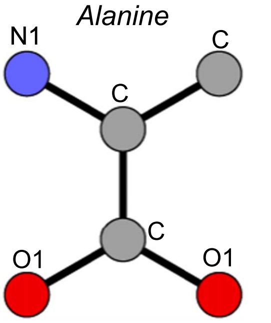


**Figure 3** Examples of C-MCS and TD-MCS for gefitinib and dacomitinib. Their IDs in the KEGG DRUG database are D01977 and D09883, respectively. Both gefitinib and dacomitinib are EGFR inhibitors. The corresponding atoms are shown in large circles. A. Connected MCS (C-MCS). The atoms in the left ring are not overlapped. B. Topologically constrained disconnected MCS (TD-MCS) with $\theta$=1. Almost all of the heavy atoms except for the linking atoms, are overlapped.



| Atom pair | Count |
|---|---|
| [C ]-[1]-[C ] | 2 |
| [O1]-[1]-[C ] | 2 |
| [N1]-[1]-[C ] | 1 |
| [C ]-[2]-[C ] | 1 |
| [O1]-[2]-[C ] | 2 |
| [O1]-[2]-[O1] | 1 |
| [N1]-[2]-[C ] | 2 |
| [O1]-[3]-[C ] | 2 |
| [N1]-[3]-[O1] | 2 |

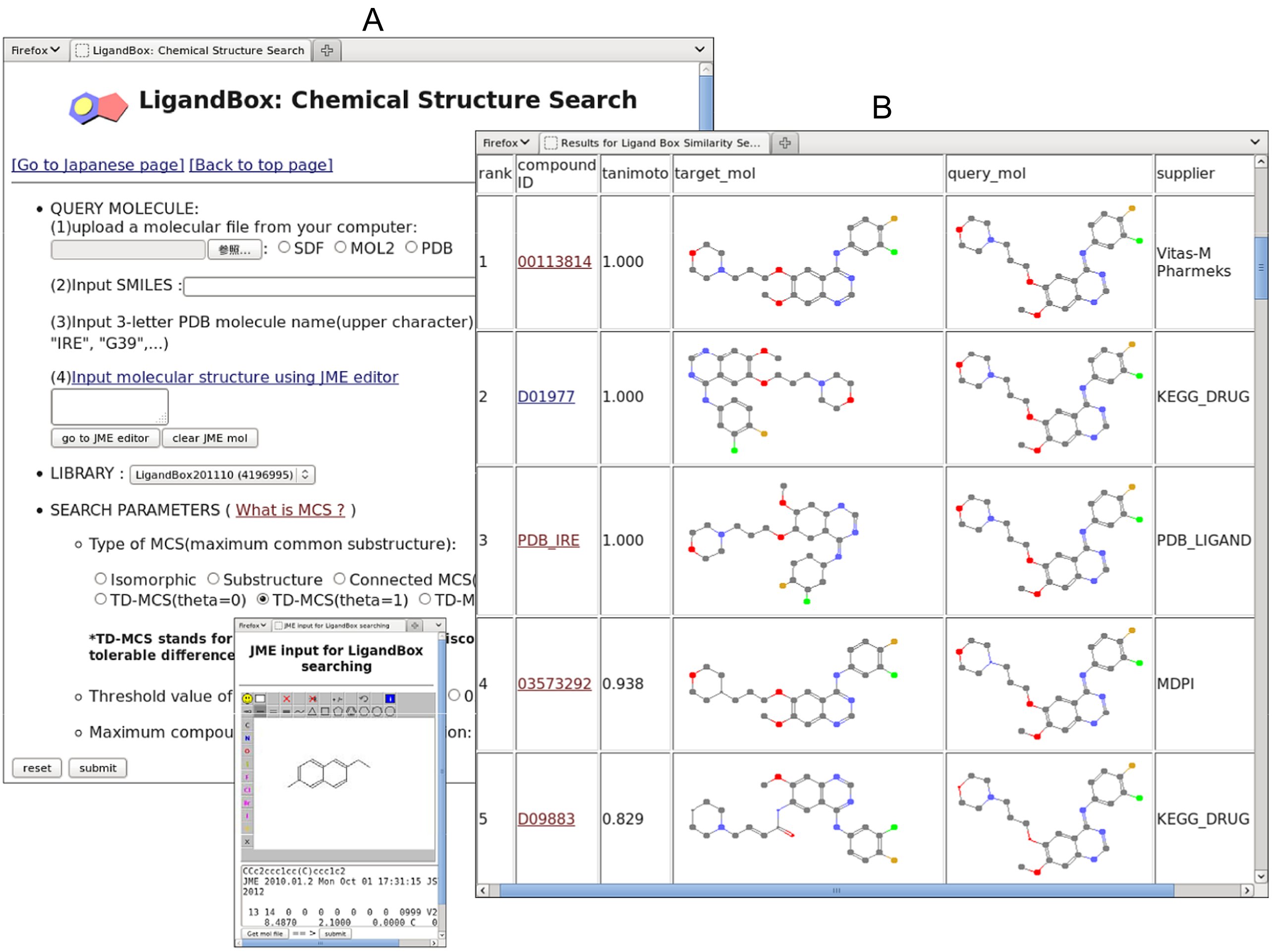
**Figure 4** An example of the atom pair descriptor proposed by Cahart *et al.*[28] for an alanine molecule. The atom pair descriptor encodes atom pairs with the atom types and the length of the shortest path distance: [atom type1] – [distance] – [atom type2]. The vector of the observed count of atom pairs is used as the descriptor. "C", "O1" and "N1" are the atom types used in the *kcombu* program in the default mode[20].

5A. A search for similar compounds against millions of compounds in the database can be accomplished within a few minutes. For the similar compounds in the library, their corresponding parts are highlighted by larger circles, as shown in Figure 5B. These similar compounds can be downloaded in a 3D MOL2 file format.

## Results and Discussions

### Comparison with the ZINC and PubChem databases

We examined the overlapping compounds with the ZINC database[13,14], which is the most popular 3D compound database, by checking for compounds with identical unique SMILES string (Table 4). This is a rough comparison, because our unique SMILES is the original classic style: it ignores all of the hydrogen atoms and the stereo chemical isomers[26,27]. "All Purchasable" compounds (updated: 2012-03-15) in the ZINC 12 database were used for the comparison. The total numbers of compounds are about 4 million for LigandBox, and 19 million for the ZINC database. The number of unique SMILES strings in ZINC is 66% of the total number, whereas that in LigandBox is not much different from the total number of compounds. This is because the ZINC database registered multiple conformers and tautomers, generated from the same 2D structure, as independent chemical entries. The number of SMILES commonly found in LigandBox and ZINC is 2.67 million, which is 64.6% of LigandBox, and 20.3% of ZINC. In the LigandBox database, 1.46 million unique compounds are stored, whereas 10 millions are unique in the ZINC database. Most of these 1.46 million unique compounds are from common suppliers working with both LigandBox and ZINC. We guess this difference may occur because versions of our compound


A
B
Firefox
LigandBox: Chemical Structure Search
Results for Ligand Box Similarity Se...
JME input for LigandBox searching
参照...
go to JME editor
clear JME mol
reset
submit
Get mol file
== >
submit

LigandBox: Chemical Structure Search

[Go to Japanese page] [Back to top page]

- QUERY MOLECULE:
  (1)upload a molecular file from your computer: ○ SDF ○ MOL2 ○ PDB
  (2)Input SMILES :
  (3)Input 3-letter PDB molecule name(upper character) "IRE", "G39",...)
  (4)Input molecular structure using JME editor
- LIBRARY : LigandBox201110 (4196995)
- SEARCH PARAMETERS ( What is MCS ? )
  - Type of MCS(maximum common substructure):
    ○ Isomorphic ○ Substructure ○ Connected MCS(
    ○ TD-MCS(theta=0) ● TD-MCS(theta=1) ○ TD-M

  *TD-MCS stands for ... isco
  tolerable difference
  - Threshold value of ... ○ 0
  - Maximum compou... on:

JME input for LigandBox searching

```
CCc2ccc1cc(C)ccc1c2
JME 2010.01.2 Mon Oct 01 17:31:15 JS
2012

 13 14  0  0  0  0  0  0  0  0999 V2
    8.4870    2.1000    0.0000 C   0
```

| rank | compound ID | tanimoto | target_mol | query_mol | supplier |
|---|---|---|---|---|---|
| 1 | 00113814 | 1.000 | | | Vitas-M<br>Pharmeks |
| 2 | D01977 | 1.000 | | | KEGG_DRUG |
| 3 | PDB_IRE | 1.000 | | | PDB_LIGAND |
| 4 | 03573292 | 0.938 | | | MDPI |
| 5 | D09883 | 0.829 | | | KEGG_DRUG |

**Figure 5** A. The chemical structure search page. B. The result page for the chemical structure search.

**Table 4** Number of common and unique compounds in LigandBox and ZINC databases

| | Total | Number of SMILES | Common | Unique |
|---|---|---|---|---|
| LigandBox | 4,196,995 | 4,127,441 | 2,667,351<br>(64.6%) | 1,460,090<br>(35.4%) |
| ZINC | 19,739,207 | 13,154,030 | 2,667,351<br>(20.3%) | 10,486,679<br>(79.7%) |

**Table 5** Number of common and unique compounds in LigandBox and PubChem databases

| | Total | Number of SMILES | Common | Unique |
|---|---|---|---|---|
| LigandBox | 4,196,995 | 4,127,441 | 2,967,406<br>(71.9%) | 1,160,035<br>(28.1%) |
| PubChem | 35,554,380 | 27,096,325 | 2,967,406<br>(11.0%) | 24,128,919<br>(89.0%) |

catalogues from the suppliers may be different from those used for the ZINC database.

Similarly, the overlapping compounds with the PubChem database were examined (Table 5). The SDF files of PubChem compounds (updated: 2012-09-06) were used for the comparison. A unique SMILES string was calculated only for the largest connected component if the compound consists of multiple connected components. The total number of PubChem compounds is about 35 million, that of unique SMILES strings in ZINC is 28 million (shown in Table 5). The number of SMILES commonly found in LigandBox and PubChem is 2.97 million, whereas that of unique SMILES compounds in LigandBox for PubChem is 1.16 million, which is not much different from number of unique compounds for ZINC (1.46 million compounds shown in Table 4). Among these 1.16 million unique compounds, 1.05 million compounds (90.7%) are common in the 1.46 unique compounds for ZINC.

For a comparison of generated 3D structures in LigandBox and ZINC, we calculated the root mean square deviations (RMSDs) for the corresponding compound pairs with identical SMILES strings. Among the 2.7 million pairs, we only used 2.3 million pairs with identical chiralities. Only heavy atoms were used for the calculation. As shown in Fig-

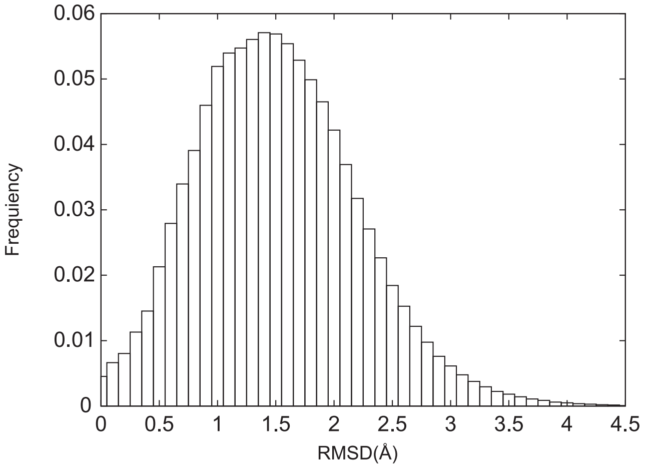


**Figure 6** Root mean square distribution (RMSD) of the corresponding compound pairs between the LigandBox and ZINC database. Corresponding compounds have the identical SMILES string and identical chiralities. Number of the corresponding compound pairs is 2,253,967.

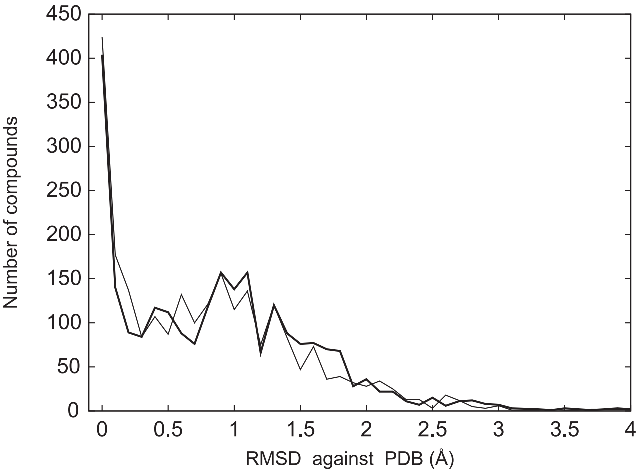


**Figure 7** Distribution of root mean square deviations (RMSDs) of the corresponding compound pairs between the LigandBox and PDB databases and those between the ZINC and PDB databases. The thick line corresponds to the RMSDs between LigandBox and PDB, whereas the thin line corresponds to the RMSDs between ZINC and PDB. The corresponding compounds have identical SMILES string and identical chiralities. Number of the corresponding compound pairs is 2,821.

ure 6, the RMSD distribution has small values: the averaged RMSD value is 1.57 Å, and the RMSDs of 79% of the pairs are less than 2.0 Å. This means that the 3D structures in LigandBox and the ZINC database are reasonably similar with some exceptions, where different stable conformers may have been chosen.

### 3D structural comparison with PDB database

We compared 3D structures of LigandBox and ZINC database with corresponding 3D structures stored in PDB to evaluate accuracies of the 3D structures. First, we selected 2,821 chemical compounds registered both in LigandBox and ZINC databases, which have corresponding compounds in PDB with the same SMILES strings and chilarities. 3D structures of compounds in PDB were obtained from the file "all-pdb.tar.gz" on the Ligand Expo database (http://ligand-expo.rcsb.org). This file contains all the 3D structures in PDB for each type of molecules. Some molecules have several 3D conformations. For example, "IRE" molecule ("iressa") has 4 structures, whereas "ATP" molecule has 1,550 structures. If more than one 3D structure are available in PDB, the smallest RMSD value was employed for the statistical analysis. Figure 7 shows RMSD distribution of LigandBox and ZINC structures with PDB structures. These two distributions are very similar. It suggests that our *myPresto* program package has a similar prediction ability to commercial programs employed by the ZINC database.

### Chiralities and protonations

When we generated the 3D conformations from the original 2D coordinates, we paid attention to their chiralities and protonation states, which may affect the accuracy of virtual screening. First, among 4 million compounds in LigandBox, 1.2 million compounds have chiral atoms. Among these chiral compounds, only 0.1 million compounds (mainly natural compounds and approved drugs) have stereo parity information in the corresponding 2D mol/SDF files from the suppliers' catalogue. For the 1.1 million compounds with ambiguous chiralities, LigandBox generated multiple 3D conformers and stored the energetically stable conformations, although this procedure increased the size and the redundancy of the database. Even if the stereo information is available, it sometimes has inconsistencies, especially when the compounds have complicated chiral ring systems. The chiral ambiguity in the database is one of the reasons why we employed the non-chiral SMILES and MCS in LigandBox.

Second, we assume the protonation of the dominant ion form at pH 7; however, this may not be realistic for all cases. Recently, the problem of tautomerism (the intramolecular movement of hydrogen from one atom to another) has been emphasized[30]. Between some tautomeric molecules, the type of a corresponding bond (such as single, double, triple) may not be the same. This means that the ambiguity of protonation can affect not only a structure-based virtual screening, but also a 2D-ligand-similarity search. For example, when we create links between different chemical databases using the MCS and SMILES comparisons, these methods may not find identical compound pairs if they have different tautomeric chemical 2D structures. We are planning to prepare multiple protonated and tautomeric forms for one molecule in the future. To find all of the tautomers of the query molecule, our chemical structure search engine (shown in Figure 5) ignores types of bonds and hydrogen atoms in the default setting.

## Conclusion

LigandBox is a 3D chemical compound library available for efficient structure-based virtual screening with a MCS structure search engine. Some of compounds in LigandBox are unique; they are not registered in ZINC and PubChem, and thus LigandBox is complementary to other related 3D chemical databases. The MCS search engine enables users to find novel similar molecules for known active molecules. We hope our database contributes to toward the discovery of new active chemical compounds by virtual screening.

## Acknowledgement

This work was supported by the New Energy and Industrial Technology Development Organization (NEDO) in Japan.

## References

1. Williams, A. J. A perspective of publicly accessible/open-access chemistry database. *Drug Discovery Today* **13**, 495–501 (2008).
2. Noorden, R. V. Chemistry's web of data expands. *Nature* **483**, 524 (2012).
3. 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.* **37**, W623–W633 (2009).
4. 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.* **40**, D1100–D1107 (2011).
5. Shoichet, B.K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).
6. Kolb, P., Ferreira, R. S., Irwin, J. J. & Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Cur. Opn. Biotech.* **20**, 429–436 (2009).
7. Cosconati, S., Forli, S., Perryman, A. L., Harris, R., Goodsell, D. S. & Olson, A. J. Virtual screening with AutoDock: theory and practice. *Expert. Opin. Drug Discov.* **5**, 597–607 (2010).
8. Taboureau, O., Baell, J. B., Fernandez-Recio, J. & Villoutreix, B. O. Established and emerging trends in computational drug discovery in the structural genomics era. *Chem. Biol.* **19**, 29–41 (2012).
9. 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* **2**, 64–72 (2009).
10. Lang, P. T., Brozell, S. R., Mukherjee, S., Pettersen, E. F., Meng, E. C., Thomas, V., Rizzo, R. C., Case, D. A., James, T. L. & Kuntz, I. D. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA* **15**, 1219–1230 (2009).
11. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. & Olson, A. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* **30**, 2785–2791 (2009).
12. Fukunishi, Y., Mikami, Y. & Nakamura, H. Similarities among receptor pockets and among compounds: Analysis and application to in silico ligand screening. *J. Mol. Graph. Model.* **24**, 34–35 (2005).
13. Irwin, J. I. & Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Info. Model.* **45**, 177–182 (2005).
14. 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.* **52**, 1757–1768 (2012).
15. Masciocchi, J., Frau, G., Fanton, M., Sturlese, M., Floris, M., Pireddu, L., Palla, P., Cedrati, F., Rodriguez-Tomé, P. & Moro, S. MMsINC: A large-scale chemoinformatics database. *Nucleic Acids Res.* **37**, D284–D290 (2009).
16. Del Rio, A., Barbosa, A. J. M., Caporuscio, F. & Mangiatordi, G.F. CoCoCo: A free suite of multiconformational chemical databases for high-throughput virtual screening purposes. *Mol. BioSyst.* **6**, 2122–2128 (2010).
17. Bolton, E. E., Chen, J., Kim, S., Han, L., He, S., Shi, W., Simonyan, V., Sun, Y., Thiessen, P. A., Wang, J., Yu, B., Zhang, J. & Bryant, S. H. PubChem3D: A new resource for scientists. *J. Cheminform.* **3**, 32 (2011).
18. Sadowski, J. & Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **93**, 2567–2581 (1993).
19. Hawkins, P. C. D., Skillman, G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structure database. *J. Chem. Info. Model.* **50**, 572–584 (2010).
20. Kawabata, T. Build-up algorithm for atomic correspondence between chemical structures. *J. Chem. Info. Model.* **51**, 1775–1787 (2011).
21. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **34**, D109–D114 (2012).
22. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Developing and testing of a general amber force field. *J. Comp. Chem.* **25**, 1157–1174 (2004).
23. Fukunishi, Y. & Nakamura, H. Definition of drug-likeness for compound affinity. *J. Chem. Info. Model.* **51**, 1012–1016 (2011).
24. Mashimo, T., Fukunishi, Y., Orita, M., Katayama, N., Fujita, S. & Nakamura, H. Quantitative analysis of aggregation-solubility relationship by in-silico solubility prediction. *International J. High Throughput Screening* **1**, 99–107 (2010).
25. Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., Kengaku, Y., Cho, H., Standley, D. M., Nakagawa, A. & Nakamura, H. Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. *Nucleic Acids Res.* **40**, D453–D460 (2011).
26. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. 2. *J. Chem. Inf.Comput. Sci.* **28**, 31–36 (1988).
27. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. a lgorithms for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101 (1989).
28. Carhart, R. E., Smith, H. S. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
29. Ertl, P. Molecular structure input on the web. *J. Cheminfo.* **2**, 1 (2010).
30. Martin, Y. C. Let's not forget tautomers. *J. Comput. Aided. Mol. Des.* **23**, 693–704 (2009).