

# CFam: a chemical families database based on iterative selection of functional seeds and seed-directed compound clustering

Cheng Zhang<sup>1,2,3</sup>, Lin Tao<sup>1,4,\*</sup>, Chu Qin<sup>1,4</sup>, Peng Zhang<sup>1</sup>, Shangying Chen<sup>1</sup>, Xian Zeng<sup>1</sup>, Feng Xu<sup>5,6</sup>, Zhe Chen<sup>6</sup>, Sheng Yong Yang<sup>2</sup> and Yu Zong Chen<sup>1,2,\*</sup>

<sup>1</sup>Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore, Singapore 117543, <sup>2</sup>State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu 610041, China, <sup>3</sup>Computational and Systems Biology, Singapore-MIT Alliance, National University of Singapore, Singapore, <sup>4</sup>NUS Graduate School for Integrative Sciences and Engineering, Singapore 117456, <sup>5</sup>College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Tianjin 300071, China, <sup>6</sup>State Key Laboratory of Medicinal Chemistry & Biology, Tianjin International Joint Academy of Biotechnology & Medicine, Tianjin 300457, China and <sup>7</sup>Zhejiang Key Laboratory of Gastro-intestinal Pathophysiology, Zhejiang Hospital of Traditional Chinese Medicine, Zhejiang Chinese Medical University, No. 54 Youdian Road, Hangzhou 310006, China

Received August 29, 2014; Accepted November 06, 2014

## ABSTRACT

Similarity-based clustering and classification of compounds enable the search of drug leads and the structural and chemogenomic studies for facilitating chemical, biomedical, agricultural, material and other industrial applications. A database that organizes compounds into similarity-based as well as scaffold-based and property-based families is useful for facilitating these tasks. CFam Chemical Family database <http://bidd2.cse.nus.edu.sg/cfam> was developed to hierarchically cluster drugs, bioactive molecules, human metabolites, natural products, patented agents and other molecules into functional families, superfamilies and classes of structurally similar compounds based on the literature-reported high, intermediate and remote similarity measures. The compounds were represented by molecular fingerprint and molecular similarity was measured by Tanimoto coefficient. The functional seeds of CFam families were from hierarchically clustered drugs, bioactive molecules, human metabolites, natural products, patented agents, respectively, which were used to characterize families and cluster compounds into families, superfamilies and classes. CFam currently contains 11 643 classes, 34 880 superfamilies and 87 136 families of 490 279 compounds (1691 approved drugs, 1228 clinical trial drugs, 12 386 inves-

tigative drugs, 262 881 highly active molecules, 15 055 human metabolites, 80 255 ZINC-processed natural products and 116 783 patented agents). Efforts will be made to further expand CFam database and add more functional categories and families based on other types of molecular representations.

## INTRODUCTION

Similarity-based clustering and classification of compounds have been extensively used in diverse tasks ranging from the search of bioactive agents for drug discovery (1–4) to the molecular and chemogenomic studies in such applications as chemspace navigation and analysis (5,6), structure-target relationship investigation (7–12), cross-pharmacology profiling of intra-family and cross-family targets (13,14) and receptor de-orphanization (15). For facilitating these and other tasks and for the orderly management of known compounds and the study of new compounds, it would be advantageous to organize the known compounds into chemical families based on structural similarity (16,17) as well as molecular scaffold classification (5,18,19) and molecular descriptor projection (19,20). This requires a method and resource for defining, generating and maintaining a comprehensive set of chemical families. To the best of our knowledge, such a resource is not yet publically available. We therefore developed the CFam Chemical Family database (<http://bidd2.cse.nus.edu.sg/cfam>) both as a database of function-based chemical families and as a re-

\*To whom correspondence should be addressed. Tel: +65 6516 6877; Fax: +65 6774 6756; Email: yzchen@cz3.nus.edu.sg  
Correspondence may also be addressed to L. Tao. Tel +65 6516 6877; Fax: +65 6774 6756; Email: linntao@hotmail.com

Search CFam by Molecule, Family, Superfamily or Class Name/ID

Search CFAM by molecule, family, superfamily or Class names or IDs.

Click [here](#) for examples.

Browse CFam [Family](#) / [Superfamily](#) / [Class](#) by Functional Category

<a href="#">Approved Drug Families</a>	<a href="#">Clinical Trial Drug Families</a>	<a href="#">Investigative Drug Families</a>	<a href="#">Bioactive Molecule Families</a>
<a href="#">Human Metabolite Families</a>	<a href="#">Natural Product Families</a>	<a href="#">Patented Agent Families</a>	<a href="#">Food Ingredient &amp; Additive Families</a>
<a href="#">Flavor &amp; Scent Families</a>	<a href="#">Agrochemical Families</a>	<a href="#">Toxic Substance Families</a>	<a href="#">Other Compound Families</a>

Align Your Molecule to CFam Families by Using [SMILES](#) or [Fingerprints](#)

Search with structure of your molecule against the CFAM database based on structural similarity.

[Click here for sample SMILES](#)

Download

A flat file containing CFAM seed information can be downloaded [here](#).

**Figure 1.** CFam web interface. CFam is searchable by three modes: compound and family name and ID searching, browsing of CFam families, superfamilies and classes and the alignment of a compound against CFam families.

source for facilitating further development of chemical family databases.

Generating a chemical family database would rely heavily on automated algorithms for classifying large number of known compounds that exceed 30 million compounds, 1.4 million bioactive molecules and 760 000 patented agents in the Pubchem (21) and ChEMBL (22) databases, which evokes two problems. One is the difficulty to strictly use hierarchical clustering algorithm for grouping such a large number of known compounds, even though k-means hierarchical clustering algorithm is capable of clustering 800 000 compounds (2,16) and non-hierarchical ones can cluster millions of compounds (23). The second is the difficulty to systematically define chemical families and select family members relevant to both structural and chemical studies and applications in pharmaceutical, biomedical, agricultural and industrial research and development. These problems also arise in generating protein domain families, which have been resolved by selecting subsets of proteins of known functions as the seeds of protein domain families to both define each family's functional and structural characteristics and select family members by multiple sequence alignment against the seed proteins (24). We employed a similar strategy for generating the CFam chemical families.

To make CFam chemical families more relevant to the applications in pharmaceutical, biomedical, agricultural, ma-

terial and other industrial applications as well as to the research in chemistry and related scientific disciplines, the seeds of the CFam families were or are to be iteratively selected from hierarchically clustered approved drugs, clinical trial drugs, investigative drugs, bioactive molecules, human metabolites, food ingredients and additives, flavors and scents, agrochemicals, natural products, patented agents, toxic substances, purchasable compounds and other known compounds based on the literature-reported high-similarity measures (25–28). These families were further clustered into CFam superfamilies and classes by hierarchically clustering the seeds based on the literature-reported intermediate similarity (11,29,30) and remote similarity (3,13,30) measures. Although this iterative hierarchical clustering procedure seems similar to the incremental clustering algorithm used in selecting representative proteins for clustering proteins (31) and representative compounds for clustering large compound libraries (23), there are two significant differences. One is that the seed selection and clustering processes are based on hierarchical clustering algorithms. The second is the preferential selection of compounds of higher functional importance as the seeds in the order of drugs, bioactive molecules, human metabolites, etc.

Currently, CFam database includes the seeds, members and names of families, superfamilies and classes functionally characterized by the approved drugs, clinical trial drugs,

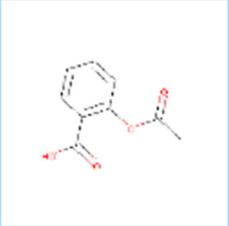
Search Results			
CFAM Mol ID	CFAMM00072836		
Family	<a href="#">CFFAD534 Cyclooxygenase inhibitor salicylate derivative Aspirin family</a>		
Superfamily	<a href="#">CFSAD463 Cyclooxygenase inhibitor salicylate derivative Aspirin Superfamily</a>		
Class	<a href="#">CFCAD426 Class 426</a>		
External ID	<a href="#">DAP000843</a>	External Source	<a href="#">TTD</a>
PubChem CID	<a href="#">2244</a>	Functional Type	Approved Drug
Molecule Name	Aspirin		
IUPAC Name	2-acetyloxybenzoic acid		
Synonyms	ACETYLSALICYLIC ACID;2-Acetoxybenzoic acid;Ecotrin;Acenterine;Acylpyrin;Polopiryna;Easprin;Acetylsalicylate;2-(Acetyloxy)benzoic acid;Acetophen		
InChi	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)		
InChiKey	BSYNRYMUTXBXSQ-UHFFFAOYSA-N		
Formula	C9H8O4	Molecular Weight	180.16
Cross Link	<a href="#">CHEMBL25</a> <a href="#">DB00945</a>		
Structure			

Figure 2. A CFam molecule page resulting from the name search by inputting 'aspirin' and selecting 'molecule'.

showing 1 to 500 of 1114.	<a href="#">&lt;&lt;First</a> <a href="#">&lt;Prev</a> <a href="#">Page 1</a> <a href="#">Next</a> <a href="#">Last&gt;&gt;</a>	
show families with	<input checked="" type="radio"/> seed and member	<input type="radio"/> seed only
<b>Browsing Family of Functional Category "Approved Drug"</b>		
Name	Seeds	Members
<a href="#">D(2) dopamine receptor ligand dibenzothiazepinone derivative Quetiapine ...</a>	2	7
<a href="#">cAMP-specific 3',5'-cyclic phosphodiesterase 4A inhibitor xanthine deriv...</a>	18	210
<a href="#">D(2) dopamine receptor ligand Benzisoxazole Derivative Risperidone family</a>	3	32
<a href="#">Abelson tyrosine-protein kinase 2 inhibitor thiazole derivative Dasatini...</a>	19	50
<a href="#">Receptor-type tyrosine-protein kinase FLT3 inhibitor indoline derivative...</a>	4	151
<a href="#">Serine/threonine-protein kinase B-raf inhibitor diaryl urea analog Soraf...</a>	12	93
<a href="#">Glucocorticoid receptor ligand glucocorticoid analog Dexamethasone Family</a>	13	248
<a href="#">5-hydroxytryptamine receptor 1B ligand triptan Almotriptan Family</a>	8	438

Figure 3. The CFam approved drug families browsing page resulting from the clicking of 'Family' in the section header titled 'Browse CFam Family/Superfamily/Class by Functional Category' and 'Approved Drug Families' in the section.

Top 30 Matches in the CFAM Database	
<b>matched families with high similarity: 5</b>	
Category	Approved Drug with Human Metabolite & Natural Product
Family	<a href="#">CFFAD534 Cyclooxygenase inhibitor salicylate derivative Aspirin family</a>
Superfamily	<a href="#">CFSAD463 Cyclooxygenase inhibitor salicylate derivative Aspirin Superfamily</a>
Class	<a href="#">CFCAD426 Class 426</a>
Category	Patent Compound
Family	<a href="#">CFFPA44136 Putative Patent EP0920416,W09807705 Family 2</a>
Superfamily	<a href="#">CFSID627 Proto-oncogene tyrosine-protein kinase Src inhibitor RU78300 Superfamily</a>
Class	<a href="#">CFCID441 Class 1791</a>
<b>matched families with intermediate similarity: 25</b>	
Category	Patent Compound
Family	<a href="#">CFFPA40333 Putative Patent EP0060640,US4822804,US4564620 Family 4</a>
Superfamily	<a href="#">CFSNP598 (6-bromo-5,8-dioxonaphthalen-1-yl) acetate 6-bromo-5,8-dioxo-5,8-dihydronaphthalen-1-yl Acetate Superfamily</a>
Class	<a href="#">CFCID875 Class 2225</a>
Category	Patent Compound
Family	<a href="#">CFFPA10344 Putative Selective Inhibitors Of Benzylaminoxidases With Respect To Other Aminoxidases In Patent US4888283,EP0210140,EP0462800 Family</a>
Superfamily	<a href="#">CFSID627 Proto-oncogene tyrosine-protein kinase Src inhibitor RU78300 Superfamily</a>
Class	<a href="#">CFCID441 Class 1791</a>

**Figure 4.** The CFam result page of the alignment of aspirin with CFam seeds.

investigative drugs, highly active molecules ( $IC_{50}$  or  $K_i < 1 \mu M$  against molecular target), human metabolites, zinc-processed natural products and patented agents. Table 1 provides the statistics of CFam seeds, compounds, families, superfamilies and classes with respect to the seven functional categories of compounds.

#### DATA COLLECTION AND PROCESSING

Because of the high computational cost of clustering large number of compounds, the first version of CFam primarily focuses on the following seven categories of compounds of functional significance: 1691 approved drugs from TTD (32) and Drugbank (33), 1228 clinical trial drugs and 12 386 investigative drugs from TTD (32), 262 881 highly active molecules ( $IC_{50}$  or  $K_i < 1 \mu M$  against molecular target) from ChEMBL version 18 (22), 15 055 human metabolites

from HMDB (34), 80 255 ZINC-processed natural products from ZINC (35) and 116 783 patented agents from PubChem (21) databases, respectively. For database entries with multiple non-linked components, only the largest component was selected. Hydrogens were added and salt ions were removed by using Open Babel (36), duplicates were identified and removed by comparative analysis of their InChIKeys, which is a hashed version of InChI (37) designed to be nearly unique for each individual compound with a collision resistance of  $2.2 \times 10^{15}$  (38).

#### GENERATION OF CFAM FAMILIES OF HIGH SIMILARITY COMPOUNDS

Molecular similarity and analysis may be conducted from different structural, physicochemical and functional

**Table 1.** The statistics of CFam seeds, compounds, families, superfamilies and classes with respect to the seven functional categories of compounds: approved drugs, clinical trial drugs, investigative drugs, bioactives (currently highly active molecules), human metabolites, zinc-processed natural products and patented agents

Functional category	Number of seeds	Number of seeds and members	Number of families	Number of superfamilies	Number of classes
Approved Drugs	1691	95 367 (4121 HM, 19 408 NP)	1114	937	813
Clinical Trial Drugs	1168	38 981 (551 HM, 3258 NP)	863	756	537
Investigative Drugs	11 093	93 191 (4321 HM, 11 881 NP)	4226	2870	1700
Bioactives	98 523	171 162 (833 HM, 24 439 NP)	29 983	15 088	4035
Human Metabolites	5229	10 408 (5229 HM, 1820 NP)	2058	1377	709
Natural Products	19 449	20 821	4017	1517	394
Patented Agents	60 349	60 349	44 875	12 335	3455
Total	197 502	490 279	87 136	34 880	11 643

The number of members of these families from the two categories of special interests, human metabolites (HM) and natural products (NP) are also provided.

perspectives by using different types of molecular representations. These include molecular descriptors (19,20,39), molecular scaffolds (5,18,19), molecular fingerprints (3,16,17) and other molecular representations, such as chemical graphs, pharmacophore patterns and molecular fields (40–43). Multiple forms of chemical families can thus be generated from these molecular representations in a similar manner as the multiple forms of protein families generated from multiple-sequence alignment of protein domains (24,44), conserved signature profiling of selected sequence segments (45), structure classification (46,47) and combined analysis of these and other features (48). Due to the high computational cost in clustering large number of compounds, in the first version of CFam, we only used one type of molecular representation, the 2D molecular fingerprints (specifically, the 881-bit PubChem substructure fingerprints computed by using PaDEL (49)), for representing molecules, which was selected because of its computational efficiency, demonstrated effectiveness in similarity searching and extensive applications in drug discovery (3,50–54). The other types of molecular representations will be used in the future version of CFam for generating other forms of chemical families.

The seeds of CFam families were assigned and used to assemble compounds into CFam families by the following iterative hierarchical clustering procedure. In the first iteration, 1691 approved drugs were clustered by hierarchical clustering algorithm with the 2D fingerprint Tarnimoto coefficient (2DF-TC) as the similarity metric and the complete linkage as the linkage criterion. Tarnimoto coefficient was used because it is the most popular similarity metric for measuring compound similarity (3). Complete linkage was used because of its relatively good performance in clustering bioactive compounds in a recent comparative study (55). The criterion for grouping compounds into a cluster of high-similarity compounds is 2DF-TC > 0.85, which was adopted because it is a widely used criterion for avoiding structural redundancy in selecting compound libraries for screening bioactive compounds (25,26). High-similarity compounds grouped by this criterion typically have 30–81% chance of having the same activity in the same bioassay (26–28). The drug/drugs in each cluster was/were assigned as the seed/seeds of a CFam-approved drug family with the family name systematically characterized by the target/targets, activity type (e.g. inhibitor),

molecular class/classes (e.g. benzisoxazole derivative) and drug name/names of the seed/seeds.

In the second iteration, the 2DF-TCs of the 1228 clinical trial drugs against the seed/seeds of the existing CFam families were first computed. If the 2DF-TC of a drug is > 0.85 with respect to all the seeds/seed of a family, the drug was assigned as a seed of that family. If the 2DF-TC of a drug is > 0.85 to some but not all of the seeds of a family, the drug was assigned as a member of that family. If the 2DF-TC of a drug is > 0.85 to the seeds of more than one family, the drug was tentatively assigned to the family/families with the largest 2DF-TC and the remaining family/families was/were marked as a cousin family to the assigned family/families and these cousins are indicated in the CFam database (e.g. CFFAD942 Prostaglandin G/H synthase 2 inhibitor diarylsubstituted isoxazole derivative valdecoxib family is a cousin family of CFFAD3 D2 dopamine receptor ligand benzisoxazole derivative risperidone family) so that the cousin families can be subsequently evaluated for possible merger into a combined family. The remaining unassigned clinical trial drugs were subject to the same procedure as that of the first iteration to assign them as the seed/seeds of CFam clinical trial drug families for assembling compounds into the respective families.

In the subsequent iterations, each set of 12 386 investigative drugs, 262 881 highly active molecules, 15 055 human metabolites, 80 255 ZINC-processed natural products and 116 783 patented agents were in turn subject to the same procedure as that of the second iteration to assign compounds into the existing CFam families or as the seed/seeds of the new CFam investigative drug families, bioactive molecule families, human metabolite families, natural product families and patented agent families for assembling compounds into the corresponding families, respectively. If the 2DF-TC of a compound is > 0.85 to the seeds of more than one family, it was preferentially assigned in order of priority to approved drug, clinical trial drug, bioactive molecule (currently highly active molecule), human metabolite, natural product and patented agent family, respectively. Certain functional categories, such as human metabolites and natural products, are of special interests beyond one scientific discipline. Therefore, if a compound from these categories (e.g. a natural product) was preferentially assigned to a family of a different category (e.g. approved drug), that family was marked and is displayed as a

family containing compound/compounds from this special category (e.g. approved drug family with natural product).

While possible, the names of these families were systematically determined in a similar manner as those of approved drugs. Many clinical trial and investigative drugs have little molecular class information and large number of bioactive compounds and natural products are without a common name, which make it difficult to automatically search for their molecular class names. Therefore, while possible, the IUPAC systematic names were used to extract common substructure names as putative molecular class names. Efforts will be made to determine the molecular classes of these families from the structure information of their seed/seeds. For the remaining families that we were unable to obtain molecular class information, their family names were tentatively characterized by the name/names or ID/IDs of their seed/seeds.

### GENERATION OF CFam SUPERFAMILIES OF INTERMEDIATE TO HIGH SIMILARITY COMPOUNDS, AND CFam CLASSES OF REMOTE TO HIGH SIMILARITY COMPOUNDS

The centroid seeds of the CFam families were further clustered by hierarchical clustering algorithm with the 2DF-TC as the similarity metric and the complete linkage as the linkage criterion, so that the CFam families can be assembled into CFam superfamilies and classes. The criterion for assembling CFam family/families into a superfamily of intermediate to high similarity compounds is 2DF-TC > 0.70, which was applied because compounds satisfying this criterion have been regarded as similar to one other (30,56) and those with slightly lower similarity typically have remote similarity (29). Compounds grouped by this intermediate-similarity criterion may have up to 30% chance of having the same activity in the same bioassay (11). These superfamilies were systematically named from the common target classes, chemical classes and individual family names of the constituent family names. A superfamily is typically composed of compounds of the same or highly similar molecular scaffolds targeting the same target, members of the same target subfamilies or target sites accommodating similar molecular scaffolds. For instance, the CFSAD2 cAMP-specific 3', 5'-cyclic phosphodiesterase, TNF inhibitor xanthine derivative superfamily includes two families of xanthine derivatives against the two targets and three families of structurally similar purine derivatives, N-alkylguanine acyclonucleosides and theobromines.

The criterion for further assembling CFam superfamily/superfamilies into CFam classes of remote to intermediate similarity compounds is 2DF-TC > 0.57, which was used because it can reasonably capture similarity compounds with cross-pharmacology relationships but not necessarily have the same activity (13). A CFam class typically consists of a large number of compounds that bind to multiple members of a target family/subfamily and/or target families/subfamilies with binding-sites accommodating similar molecular scaffolds, which makes it difficult to systematically name it. Therefore, CFam classes were tentatively named by their CFam class IDs only. Efforts will be made to manually determine their

names. An example of a CFam class is CFCAD3, which is composed of the binders of GPCR Class A subfamilies A1 (C-C chemokine receptors), A9 (neuropeptide Y receptors), A13 (cannabinoid receptors), A17 (dopamine receptors), A18 (muscarinic acetylcholine receptors) and A19 (5-HT receptors), cholinesterases, tryptases, dopamine transporters and sodium channel proteins, etc.

### DATABASE STRUCTURE AND ACCESS

CFam can be searched by three different modes (Figure 1). The first mode enables the search of CFam by inputting a compound name or ID (currently support CFam, Pubchem, ChEMBL, Zinc and TTD compound IDs), a CFam family name or ID, a CFam superfamily name or ID and a CFam Class ID, respectively. The relevant information may be obtained by clicking the buttons of 'Molecule', 'Family', 'Superfamily' and 'Class', respectively. For instance, inputting 'aspirin' and then clicking 'Molecule' leads to the CFam molecule CFAMM00072836 page which shows that aspirin belongs to the CFam CFFAD534 cyclooxygenase inhibitor salicylate derivative aspirin family (Figure 2). The second mode enables the browsing of CFam families, superfamilies and classes of any functional category, respectively, which can be proceeded by first clicking the 'Family', 'Superfamily' or 'Class' word in the section header titled 'Browse CFam Family/Superfamily/Class by Functional Category', and then clicking a specific functional category below the header. For instance, clicking 'Family' and then 'Approved Drug Families' leads to the page of CFam approved drug families list (Figure 3). The third mode facilitates the alignment of an input compound in SMILES or molecular fingerprint format against CFam seeds to identify CFam families with high, intermediate and remote similarity to the input compound. The list of up to 30 CFam families with at least one seed having 2DF-TC > 0.85 (high similarity family),  $0.85 \geq 2DF-TC > 0.7$  (intermediate similarity family) and  $0.7 \geq 2DF-TC > 0.57$  (remote similarity) to the input compound is provided. Figure 4 shows the result page of the alignment of aspirin with CFam seeds. To facilitate the development of chemical family databases and the structural and functional analysis of molecules, CFam seeds can be downloaded from the CFam main page (Figure 1).

### REMARKS

Specialized chemical information resources, such as the chemical family databases, complement the general chemical databases for facilitating focused studies on the navigation, classification and the structural and functional characterization of molecules. The chemical family databases that comprehensively cover the known chemspace and characterize molecules from different molecular representations are increasingly needed given the rapidly expanding pools of molecules from synthetic and natural sources (57–59) and the increasing need to analyze higher number and more variety of compounds for diverse applications (13–15,19). To meet such a need, CFam will be further updated to expand existing functional families and add new families of moderately active molecules (IC<sub>50</sub> or Ki 1–10 μM against

molecular target), food ingredients and additives, flavors and scents, agrochemicals, natural products beyond ZINC processed ones, toxic substances, purchasable compounds and other compounds. Although some of the CFam families are currently composed of seeds only, these seeds are nonetheless useful for facilitating further development of chemical families and function-based classification of compounds.

## FUNDING

Funding for open access charge: Major State Basic Research Development Program of China [2013CB967204]. The authors would also like to thank the Singapore Academic Research Fund (R-148-000-181-112).

*Conflict of interest statement.* None declared.

## REFERENCES

1. Gruneberg, S., Stubbs, M.T. and Klebe, G. (2002) Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.*, **45**, 3588–3602.
2. Bocker, A., Schneider, G. and Teckentrup, A. (2006) NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model*, **46**, 2220–2229.
3. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.
4. Riniker, S., Fechner, N. and Landrum, G.A. (2013) Heterogeneous classifier fusion for ligand-based virtual screening: or, how decision making by committee can be a good thing. *J. Chem. Inf. Model*, **53**, 2829–2836.
5. Lipinski, C. and Hopkins, A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.
6. Renner, S., van Otterlo, W.A., Dominguez Seoane, M., Mocklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T.I., Steinhilber, D. *et al.* (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.*, **5**, 585–592.
7. Hu, Y. and Bajorath, J. (2012) Rationalizing structure and target relationships between current drugs. *AAPS J.*, **14**, 764–771.
8. Eckert, H. and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, **12**, 225–233.
9. Wang, Y. and Bajorath, J. (2009) Development of a compound class-directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching. *J. Chem. Inf. Model*, **49**, 1369–1376.
10. Vogt, I., Ahmed, H.E., Auer, J. and Bajorath, J. (2008) Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Divers*, **12**, 25–40.
11. Biniashvili, T., Schreiber, E. and Klinger, Y. (2012) Improving classical substructure-based virtual screening to handle extrapolation challenges. *J. Chem. Inf. Model*, **52**, 678–685.
12. Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G. and Tang, Y. (2012) Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *J. Chem. Inf. Model*, **52**, 1103–1113.
13. Brioso, F., Carrascosa, M.C., Oprea, T.I. and Mestres, J. (2011) Cross-pharmacology analysis of G protein-coupled receptors. *Curr. Top Med. Chem.*, **11**, 1956–1963.
14. Lin, H., Sassano, M.F., Roth, B.L. and Shoichet, B.K. (2013) A pharmacological organization of G protein-coupled receptors. *Nat. Methods*, **10**, 140–146.
15. van der Horst, E., Peironcelly, J.E., Ijzerman, A.P., Beukers, M.W., Lane, J.R., van Vlijmen, H.W., Emmerich, M.T., Okuno, Y. and Bender, A. (2010) A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. *BMC Bioinformatics*, **11**, 316.
16. Bocker, A., Derksen, S., Schmidt, E., Teckentrup, A. and Schneider, G. (2005) A hierarchical clustering approach for large compound libraries. *J. Chem. Inf. Model*, **45**, 807–815.
17. Engels, M.F., Gibbs, A.C., Jaeger, E.P., Verbinnen, D., Lobanov, V.S. and Agrafiotis, D.K. (2006) A cluster-based strategy for assessing the overlap between large chemical libraries and its application to a recent acquisition. *J. Chem. Inf. Model*, **46**, 2651–2660.
18. Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T.I., Mutzel, P. and Waldmann, H. (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.*, **5**, 581–583.
19. Lachance, H., Wetzel, S., Kumar, K. and Waldmann, H. (2012) Charting, navigating, and populating natural product chemical space for drug discovery. *J. Med. Chem.*, **55**, 5989–6001.
20. Le Guilloux, V., Colliandre, L., Bourg, S., Guenegou, G., Dubois-Chevalier, J. and Morin-Allory, L. (2011) Visual characterization and diversity quantification of chemical libraries: 1. Creation of delimited reference chemical subspaces. *J. Chem. Inf. Model*, **51**, 1762–1774.
21. Bolton, E., Wang, Y., Thiessen, P.A. and SH, B. (2008) PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.*, **4**, 217–240.
22. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Kruger, F.A., Light, Y., Mak, L., McGlinchey, S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–1090.
23. Li, W. (2006) A fast clustering algorithm for analyzing highly similar compounds of very large libraries. *J. Chem. Inf. Model*, **46**, 1919–1923.
24. Sonnhammer, E.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
25. Matter, H. (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.*, **40**, 1219–1229.
26. Martin, Y.C., Kofron, J.L. and Traphagen, L.M. (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.
27. Cramer, R.D., Jilek, R.J., Guessregen, S., Clark, S.J., Wendt, B. and Clark, R.D. (2004) ‘Lead hopping’. Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem.*, **47**, 6777–6791.
28. Dunkel, M., Gunther, S., Ahmed, J., Wittig, B. and Preissner, R. (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.
29. Godden, J.W., Stahura, F.L. and Bajorath, J. (2005) Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model*, **45**, 1812–1819.
30. Boehm, M., Wu, T.Y., Claussen, H. and Lemmen, C. (2008) Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.*, **51**, 2468–2480.
31. Li, W., Jaroszewski, L. and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
32. Qin, C., Zhang, C., Zhu, F., Xu, F., Chen, S.Y., Zhang, P., Li, Y.H., Yang, S.Y., Wei, Y.Q., Tao, L. *et al.* (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.*, **42**, D1118–D1123.
33. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
34. Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E. *et al.* (2013) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–D807.
35. Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S. and Coleman, R.G. (2012) ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model*, **52**, 1757–1768.
36. O’Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T. and Hutchison, G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.
37. International Union of Pure and Applied Chemistry. (2011) InChI version 1 (software version 1.04 for Standard and Non-Standard InChI/InChIKey). <http://www.iupac.org/InChI/>
38. InChI Trust. (2012) IUPAC International Chemical Identifier (InChI) Programs InChI version 1, software version 1.04 User’s

- Guide,  
<http://www.inchi-trust.org/download/104/InChI.UserGuide.pdf>.
39. Bender, A., Jenkins, J.L., Scheiber, J., Sukuru, S.C., Glick, M. and Davies, J.W. (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model*, **49**, 108–119.
  40. Dean, P.M. (ed.). (1994) *Molecular Similarity in Drug Design*. Chapman and Hall, London.
  41. Willett, P., Barnard, J.M. and Downs, G.M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.*, **38**, 983–996.
  42. Nikolova, N. and Jaworska, J. (2003) Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.*, **22**, 1006–1026.
  43. Bender, A. and Glen, R.C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, **2**, 3204–3218.
  44. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
  45. Sigrist, C.J., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2013) New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
  46. Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
  47. Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. and Orengo, C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.
  48. Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
  49. Yap, C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.
  50. Brown, R. and Martin, Y. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.*, **37**, 1–9.
  51. Schuffenhauer, A., Gillet, V.J. and Willett, P. (2000) Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.*, **40**, 295–307.
  52. Makara, G.M. (2001) Measuring molecular similarity and diversity: total pharmacophore diversity. *J. Med. Chem.*, **44**, 3563–3571.
  53. Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today*, **7**, 903–911.
  54. Cruciani, G., Pastor, M. and Mannhold, R. (2002) Suitability of molecular descriptors for database mining. A comparative analysis. *J. Med. Chem.*, **45**, 2685–2694.
  55. Smieja, M., Warszycki, D., Tabor, J. and Bojarski, A.J. (2014) Asymmetric clustering index in a case study of 5-HT1A receptor ligands. *PLoS One*, **9**, e102069.
  56. Xue, L., Godden, J.W. and Bajorath, J. (1999) Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.*, **39**, 881–886.
  57. Thomas, G.L. and Johannes, C.W. (2011) Natural product-like synthetic libraries. *Curr. Opin. Chem. Biol.*, **15**, 516–522.
  58. Lopez-Vallejo, F., Giulianotti, M.A., Houghten, R.A. and Medina-Franco, J.L. (2012) Expanding the medically relevant chemical space with compound libraries. *Drug Discov. Today*, **17**, 718–726.
  59. van Hattum, H. and Waldmann, H. (2014) Biology-oriented synthesis: harnessing the power of evolution. *J. Am. Chem. Soc.*, **136**, 11853–11859.