DATA ARTICLE

Follow up: Compound data sets and software tools for chemoinformatics and medicinal chemistry applications: update and data transfer [v1; ref status: indexed, http://f1000r.es/32j]

Ye Hu, Jürgen Bajorath

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms University, Bonn, D-53113, Germany

v1 **First published:** 11 Mar 2014, **3**:69 (doi: 10.12688/f1000research.3713.1)
Latest published: 11 Mar 2014, **3**:69 (doi: 10.12688/f1000research.3713.1)

Original article:

Freely available compound data sets and software tools for chemoinformatics and computational medicinal chemistry applications [v1; ref status: indexed, http://f1000r.es/Mu9krs]
Ye Hu, Jurgen Bajorath

Published 14 Aug 2012

Abstract

In 2012, we reported 30 compound data sets and/or programs developed in our laboratory in a data article and made them freely available to the scientific community to support chemoinformatics and computational medicinal chemistry applications. These data sets and computational tools were provided for download from our website. Since publication of this data article, we have generated 13 new data sets with which we further extend our collection of publicly available data and tools. Due to changes in web servers and website architectures, data accessibility has recently been limited at times. Therefore, we have also transferred our data sets and tools to a public repository to ensure full and stable accessibility. To aid in data selection, we have classified the data sets according to scientific subject areas. Herein, we describe new data sets, introduce the data organization scheme, summarize the database content and provide detailed access information in ZENODO (doi: 10.5281/zenodo.8451 and doi:10.5281/zenodo.8455).

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 11 Mar 2014	✓ report	✓ report	✓ report

1 **Ajay Jain**, University of California San Francisco USA

2 **Chris J. Swain**, Cambridge Med Chem Consulting UK

3 **Patrick Walters**, Vertex Pharmaceuticals Incorporated USA

Discuss this article

Comments (1)

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

How to cite this article: Hu Y and Bajorath J. **Compound data sets and software tools for chemoinformatics and medicinal chemistry applications: update and data transfer [v1; ref status: indexed, http://f1000r.es/32j]** *F1000Research* 2014, **3**:69 (doi: 10.12688/f1000research.3713.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were declared.

First published: 11 Mar 2014, **3**:69 (doi: 10.12688/f1000research.3713.1)
First indexed: 17 Apr 2014, **3**:69 (doi: 10.12688/f1000research.3713.1)

Introduction

The compound data sets reported in our original article[1] and the new data sets presented herein have resulted from research in the chemoinformatics and medicinal chemistry area and have mostly been generated from public domain repositories of compound structures and activity data. In addition, software tools made publicly available have also been developed in our laboratory[1]. Data sets reported in the scientific literature in the context of computational method development and evaluation are often not publicly available, which limits the reproducibility of computational investigations and comparisons of different computational methods. We believe that it is important to provide such data to the scientific community to further improve the transparency and credibility of computational studies and support method development. In addition to the data sets designed for the development and evaluation of computational methods, we also make available data sets that were generated as a resource and knowledge base for medicinal chemistry applications. Our data sets and tools are provided via the ZENODO platform (https://zenodo.org/) to ensure easy and stable access.

Materials and methods

The data sets reported herein were predominantly generated from ChEMBL[2,3], BindingDB[4] and PubChem[5] (a few exceptions are specified in the original data article[1]). Compound structures are represented as SMILES[6] strings or SD files[7]. Activity information and other (data set-dependent) annotations are provided in the individual data files. For software tools (written in different languages), the source code is also made available.

Data description

Table 1 provides the updated list and classification of all freely available data sets and programs. Entries were organized according to the following scientific subject areas: data sets for structure-activity relationship (SAR) and structure-selectivity relationship (SSR) analysis, SAR visualization (SAR_VZ), and virtual screening via similarity searching or machine learning (VS_ML). In addition, the programs are provided separately (PROG). Data sets and programs are contained in separate ZENODO deposition sets with a unique reference. Three matched molecular pair (MMP)-based data sets also included in our update have recently been reported and described in detail[8]. Entries 1–30 in Table 1 represent the data sets and programs that we initially provided *via* our website[1] and entries 31–43 represent new data sets. In the following, the new data sets are described:

Entry 31

50 compound activity classes (AC) are prioritized for the evaluation of scaffold hopping potential in ligand-based virtual screening[38]. These AC contain the largest proportion of scaffold pairs with largest chemical inter-scaffold distances[38] that can be derived from current bioactive compounds and hence present challenging test cases for scaffold hopping analysis.

Entry 32

596 SAR transfer series with regular potency progression (SAR-TS-RP) are extracted from 61 AC[39]. Each SAR-TS-RP represents two compound series with different core structures and pairwise corresponding substitutions that yield comparable potency progression against a given target. These series provide a knowledge base for the analysis and prediction of SAR transfer events.

Entry 33

Four sets of molecular scaffolds (with each scaffold representing more than ten compounds) are provided that are active against a single target (ST), multiple targets from the same family (SF), or multiple targets from different families (MF)[40]. Data sets are separately assembled for different types of potency measurements (*i.e.*, K_i and IC_{50} values) and provide a resource of scaffolds representing compounds with varying degrees of target promiscuity.

Entry 34

Two multi-target compound data sets consist of confirmed screening hits[41]. Each set contains compounds with single-, dual-, and triple-target activity, or no activity. These data provide test cases for machine learning or other approaches to differentiate between compounds with overlapping yet distinct activity profiles.

Entry 35

Four multi-target compound data sets are provided[42]. Each set contains compounds tested in three different assays. Compounds are organized into eight different subsets according to their activity profiles, *i.e.*, single-, dual-, and triple-target activity, or no activity. In addition, three multi-mechanism compound sets are designed[42]. In the latter case, compounds are organized into four subsets according to their mechanism-of-action. These data sets also represent test cases for machine learning to distinguish compounds with different activity profiles or mechanisms.

Entry 36

2337 non-redundant compound series matrices (CSMs) are generated covering compounds active against a wide spectrum of targets[43]. Each matrix contains at least two analogous matching molecular series (MMS) with structurally related yet distinct cores. A matrix consists of known active compounds and structurally related virtual compounds and hence provides suggestions for compound design.

Entry 37

128 target-based data sets are assembled that consist of at least 100 compounds with precisely specified equilibrium constants (K_i values) below 1 μM for human targets[44]. These high-confidence activity data sets provide a sound basis for SAR exploration.

Entry 38

30,452 and 45,607 target-based MMS with K_i and IC_{50} values, respectively, are extracted from bioactive compounds[45].

Entry 39

221 scaffolds are identified that only occur in approved drugs but are not found in currently available bioactive compounds[46]. Accordingly, these scaffolds have been termed drug-unique scaffolds.

Entry 40

92,734 MMPs are generated from 435 AC on a basis of retrosynthetic rules[47]. These MMPs consider chemical reaction information and should be useful for practical medicinal chemistry applications.

Table 1. Data sets and programs.

Entry	Year	Subject area index label	Description
1[9]	2007	VS_ML_1	9 activity classes (AC) with increasing structural diversity
2[9]	2007	VS_ML_2	~1.44 million ZINC compounds used for various virtual screening trials
3[10]	2007	PROG_1	Molecular similarity histogram filtering
4[11]	2007	SSR_1	4 SD files with 26 selectivity sets; compounds are annotated with selectivity values for different targets
5[12]	2008	SSR_2	7 compound selectivity sets containing 267 biogenic amine GPCR antagonists
6[13]	2008	SSR_3	18 selectivity sets for targets from 4 families
7[14]	2008	VS_ML_3	25 sets of compounds of increasing complexity and size
8[15]	2009	VS_ML_4	242 hERG inhibitors
9[16]	2009	SSR_4	243 ionotropic glutamate ion channel antagonists
10[17]	2009	PROG_2	Combinatorial analog graph (CAG) program with a sample set consisting of 51 thrombin inhibitors
11[18]	2009	VS_ML_5	20 AC from the literature and 15 AC from the Molecular Drug Data Report
12[19]	2010	VS_ML_6	8 AC
13[20]	2010	PROG_3	Program to generate target selectivity patterns of scaffolds
14[21]	2010	PROG_4	Multi-target CAGs (see also entry 10) with a sample set containing 33 kinase inhibitors
15[22]	2010	PROG_5	SARANEA
16[23]	2010	PROG_6	3D activity landscape program with a sample set containing 248 cathepsin S inhibitors
17[24]	2010	SAR_1	2 sets of MMPs from BindingDB and ChEMBL
18[25]	2010	PROG_7	Similarity-potency tree (SPT) program with a sample set containing 874 factor Xa inhibitors
19[26]	2010	VS_ML_7	17 target-directed compound sets; each set contains a minimum of 10 distinct scaffolds and each scaffold represents 5 compounds
20[27]	2011	SAR_VZ	10,489 malaria screening hits
21[28]	2011	SAR_2	458 target-based sets with scaffolds and scaffold hierarchies
22[29]	2011	SAR_VZ	4 sets of compounds active against 3 or 4 targets
23[30]	2011	SAR_VZ	881 factor Xa inhibitors
24[31]	2011	VS_ML_8	50 AC prioritized for similarity searching
25[32]	2011	VS_ML_9	25 data sets from successful ligand-based virtual screening applications
26[33]	2011	SAR_3	26 conserved scaffolds in activity profile sequences of length 4
27[34]	2011	PROG_8	Scaffold distance function
28[35]	2011	SAR_4	2 sets of compounds with multiple K_i or IC_{50} measurements against the same targets that differed within 1 order of magnitude
29[36]	2012	SAR_VZ	4 AC
30[37]	2012	SAR_5	5 sets of different types of activity cliffs
31[38]	2012	VS_ML_10	50 AC for scaffold hopping analysis
32[39]	2012	SAR_6	61 AC consisting of SAR transfer series with regular potency progression
33[40]	2013	SAR_7	4 activity measurement type-dependent sets of scaffolds
34[41]	2013	VS_ML_11	2 multi-target compound sets
35[42]	2013	VS_ML_12	4 multi-target compound sets and 3 multi-mechanism sets
36[43]	2013	SAR_8	2337 compound series matrices
37[44]	2013	SAR_9	128 AC containing ≥100 compounds with K_i values
38[45]	2014	SAR_10	30,452 and 45,607 target-based MMS with K_i and IC_{50} values, respectively
39[46]	2014	SAR_11	221 drug-unique scaffolds
40[47]	2014	SAR_12	92,734 MMPs based upon retrosynthetic rules for 435 AC
41[8]	2014	SAR_13	20,073 and 25,297 MMP-based activity cliffs with K_i and IC_{50} values, respectively
42[8]	2014	SAR_14	4 activity measurement type-dependent sets of SAR transfer series with approximate or regular potency progression
43[8]	2014	SAR_15	169,889 and 240,322 transformation size-restricted MMPs based upon retrosynthetic rules with K_i and IC_{50} values, respectively

Data entries are organized according to scientific subject areas: structure-activity relationship (SAR) and structure-selectivity relationship (SSR) analysis, SAR visualization (SAR_VZ), virtual screening *via* similarity searching or machine learning (VS_ML), and programs (PROG). References in the Entry column provide the original publication introducing the program and/or data set. Program entries are described in more detail in Table 2 of our original data article[1]. The new compound data sets 31–43 are discussed in the text. Programs and data sets reported herein have been separately deposited in ZENODO for access and download.

Entry 41
20,073 and 25,297 MMP-based activity cliffs (*i.e.* pairs of structurally analogous compounds with an at least 100-fold difference in potency) are extracted from specifically active compounds based upon K_i and IC_{50} values, respectively[8]. The MMP-based activity cliffs provide a large knowledge base for SAR analysis.

Entry 42
157 and 513 MMP-based SAR transfer series with approximate potency progression plus 60 and 322 SAR transfer series with regular potency progression based upon K_i and IC_{50} values, respectively, are isolated from bioactive compounds. These transfer series are active against individual targets[8]. Similar to MMP-based activity cliffs, SAR transfer series provide a resource for SAR analysis and compound design.

Entry 43
169,889 and 240,322 transformation size-restricted MMPs based upon retrosynthetic rules with K_i and IC_{50} values, respectively, are systematically extracted from available AC[8]. Different from the retrosynthetic rule-based MMPs presented above, applied transformation size-restrictions ensure that chemical changes distinguishing compounds in pairs are small.

Summary
Herein we have provided an updated release of data sets and programs for chemoinformatics and medicinal chemistry that we make freely available. In total, 13 new data sets are introduced. Transferring all data entries in an organized form to the ZENODO platform makes them easily accessible. We hope that our current release might be of interest and helpful to many investigators in academia and the pharmaceutical industry.

Data availability
ZENODO: Programs for chemoinformatics and computational medicinal chemistry, doi: 10.5281/zenodo.8451[48].

ZENODO: Data sets for chemoinformatics and computational medicinal chemistry, doi: 10.5281/zenodo.8455[49].

Author contributions
JB designed the study, YH collected and organized the data, YH and JB wrote the manuscript.

Competing interests
No competing interests were declared.

Grant information
The author(s) declared that no grants were involved in supporting this work.

Acknowledgments
We are grateful to current and former members of our research group who have contributed to the development of the data sets and programs reported herein.

References

1. Hu Y, Bajorath J: **Freely available compound data sets and software tools for chemoinformatics and computational medicinal chemistry applications [v1; ref status: indexed, http://f1000r.es/Mu9krs].** *F1000Res.* 2012; **1**: 11. PubMed Abstract | Publisher Full Text | Free Full Text
2. Gaulton A, Bellis LJ, Bento AP, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1100–D1107. PubMed Abstract | Publisher Full Text | Free Full Text
3. Bento AP, Gaulton A, Hersey A, *et al.*: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1083–D1090. PubMed Abstract | Publisher Full Text
4. Liu T, Lin Y, Wen X, *et al.*: **BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities.** *Nucleic Acids Res.* 2007; **35**(Database issue): D198–D201. PubMed Abstract | Publisher Full Text | Free Full Text
5. Wang Y, Xiao J, Suzek TO, *et al.*: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res.* 2009; **37** (Web Server issue): W623–W633. PubMed Abstract | Publisher Full Text | Free Full Text
6. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci.* 1988; **28**(1): 31–36. Publisher Full Text
7. Dalby A, Nourse JG, Hounshell WD, *et al.*: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *J Chem Inf Comput Sci.* 1992; **32**(3): 244–255. Publisher Full Text
8. Hu Y, de la Vega de León A, Zhang B, *et al.*: **Matched molecular pair-based data sets for computer-aided medicinal chemistry [v2; ref status: indexed, http://f1000r.es/309].** *F1000Res.* 2014; **3**: 36. Publisher Full Text
9. Tovar A, Eckert H, Bajorath J: **Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity.** *ChemMedChem.* 2007; **2**(2): 208–217. PubMed Abstract | Publisher Full Text
10. Wang Y, Godden JW, Bajorath J: **A novel descriptor histogram filtering method for database mining and the identification of active molecules.** *Lett Drug Design Discov.* 2007; **4**(4): 286–292. Publisher Full Text
11. Stumpfe D, Ahmed H, Vogt I, *et al.*: **Methods for computer-aided chemical biology. Part 1: Design of a benchmark system for the evaluation of compound selectivity.** *Chem Biol Drug Des.* 2007; **70**(3): 182–194. PubMed Abstract | Publisher Full Text
12. Vogt I, Ahmed HE, Auer J, *et al.*: **Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping.** *Mol Divers.* 2008; **12**(1): 25–40. PubMed Abstract | Publisher Full Text
13. Stumpfe D, Geppert H, Bajorath J: **Methods for computer-aided chemical biology. Part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification.** *Chem Biol Drug Des.* 2008; **71**(6): 518–528. PubMed Abstract | Publisher Full Text
14. Wang Y, Geppert H, Bajorath J: **Random reduction in fingerprint bit density improves compound recall in search calculations using complex reference molecules.** *Chem Biol Drug Des.* 2008; **71**(6): 511–517. PubMed Abstract | Publisher Full Text
15. Nisius B, Göller AH, Bajorath J: **Combining cluster analysis, feature selection and multiple support vector machine models for the identification of human ether-a-go-go related gene channel blocking compounds.** *Chem Biol Drug Des.* 2009; **73**(1): 17–25. PubMed Abstract | Publisher Full Text
16. Ahmed H, Geppert H, Stumpfe D, *et al.*: **Methods for computer-aided chemical biology. Part 4: selectivity searching for ion channel ligands and mapping of molecular fragments as selectivity markers.** *Chem Biol Drug Des.* 2009; **73**(3): 273–282. PubMed Abstract | Publisher Full Text

17. Peltason L, Weskamp N, Teckentrup A, *et al.*: **Exploration of structure-activity relationship determinants in analogue series.** *J Med Chem.* 2009; **52**(10): 3212–3224.
PubMed Abstract | Publisher Full Text

18. Nisius B, Bajorath J: **Molecular fingerprint recombination: generating hybrid fingerprints for similarity searching from different fingerprint types.** *ChemMedChem.* 2009; **4**(11): 1859–1863.
PubMed Abstract | Publisher Full Text

19. Batista J, Tan L, Bajorath J: **Atom-centered interacting fragments and similarity search applications.** *J Chem Inf Model.* 2010; **50**(1): 79–86.
PubMed Abstract | Publisher Full Text

20. Hu Y, Bajorath J: **Exploring target-selectivity patterns of molecular scaffolds.** *ACS Med Chem Lett.* 2010; **1**(2): 54–58.
Publisher Full Text

21. Wassermann AM, Peltason L, Bajorath J: **Computational analysis of multi-target structure-activity relationships to derive preference orders for chemical modifications toward target selectivity.** *ChemMedChem.* 2010; **5**(6): 847–858.
PubMed Abstract | Publisher Full Text

22. Lounkine E, Wawer M, Wassermann AM, *et al.*: **SARANEA: a freely available program to mine structure-activity and structure-selectivity relationship information in compound data sets.** *J Chem Inf Model.* 2010; **50**(1): 68–78.
PubMed Abstract | Publisher Full Text

23. Peltason L, Iyer P, Bajorath J: **Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs.** *J Chem Inf Model.* 2010; **50**(6): 1021–1033.
PubMed Abstract | Publisher Full Text

24. Wassermann AM, Bajorath J: **Chemical substitutions that introduce activity cliffs across different compound classes and biological targets.** *J Chem Inf Model.* 2010; **50**(7): 1248–1256.
PubMed Abstract | Publisher Full Text

25. Wawer M, Bajorath J: **Similarity-potency trees: a method to search for SAR information in compound data sets and derive SAR rules.** *J Chem Inf Model.* 2010; **50**(8): 1395–1409.
PubMed Abstract | Publisher Full Text

26. Vogt M, Stumpfe D, Geppert H, *et al.*: **Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening.** *J Med Chem.* 2010; **53**(15): 5707–5715.
PubMed Abstract | Publisher Full Text

27. Wawer M, Bajorath J: **Extracting SAR information from a large collection of anti-malarial screening hits by NSG-SPT analysis.** *ACS Med Chem Lett.* 2011; **2**(3): 201–206.
Publisher Full Text

28. Hu Y, Bajorath J: **Combining horizontal and vertical substructure relationships in scaffold hierarchies for activity prediction.** *J Chem Inf Model.* 2011; **51**(2): 248–257.
PubMed Abstract | Publisher Full Text

29. Dimova D, Wawer M, Wassermann AM, *et al.*: **Design of multitarget activity landscapes that capture hierarchical activity cliff distributions.** *J Chem Inf Model.* 2011; **51**(2): 258–266.
PubMed Abstract | Publisher Full Text

30. Wawer M, Bajorath J: **Local structural changes, global data views: graphical substructure-activity relationship trailing.** *J Med Chem.* 2011; **54**(8): 2944–2951.
PubMed Abstract | Publisher Full Text

31. Heikamp K, Bajorath J: **Large-scale similarity search profiling of ChEMBL compound data sets.** *J Chem Inf Model.* 2011; **51**(8): 1831–1839.
PubMed Abstract | Publisher Full Text

32. Ripphausen P, Wassermann AM, Bajorath J: **REPROVIS-DB: a benchmark system for ligand-based virtual screening derived from reproducible prospective applications.** *J Chem Inf Model.* 2011; **51**(10): 2467–2473.
PubMed Abstract | Publisher Full Text

33. Hu Y, Bajorath J: **Activity profile sequences: a concept to account for the progression of compound activity in target space and to extract SAR information from analogue series with multiple target annotations.** *ChemMedChem.* 2011; **6**(12): 2150–2154.
PubMed Abstract | Publisher Full Text

34. Li R, Stumpfe D, Vogt M, *et al.*: **Development of a method to consistently quantify the structural distance between scaffolds and to assess scaffold hopping potential.** *J Chem Inf Model.* 2011; **51**(10): 2507–2514.
PubMed Abstract | Publisher Full Text

35. Stumpfe D, Bajorath J: **Assessing the confidence level of public domain compound activity data and the impact of alternative potency measurements on SAR analysis.** *J Chem Inf Model.* 2011; **51**(12): 3131–3137.
PubMed Abstract | Publisher Full Text

36. Gupta-Ostermann D, Hu Y, Bajorath J: **Introducing the LASSO graph for compound data set representation and structure-activity relationship analysis.** *J Med Chem.* 2012; **55**(11): 5546–5553.
PubMed Abstract | Publisher Full Text

37. Hu Y, Bajorath J: **Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database.** *J Chem Inf Model.* 2012; **52**(7): 1806–1811.
PubMed Abstract | Publisher Full Text

38. Li R, Bajorath J: **Systematic assessment of scaffold distances in ChEMBL: prioritization of compound data sets for scaffold hopping analysis in virtual screening.** *J Comput Aided Mol Des.* 2012; **26**(10): 1101–1109.
PubMed Abstract | Publisher Full Text

39. Zhang B, Wassermann AM, Vogt M, *et al.*: **Systematic assessment of compound series with SAR transfer potential.** *J Chem Inf Model.* 2012; **52**(12): 3138–3143.
PubMed Abstract | Publisher Full Text

40. Hu Y, Bajorath J: **Systematic identification of scaffolds representing compounds active against individual targets and single or multiple target families.** *J Chem Inf Model.* 2013; **53**(2): 312–326.
PubMed Abstract | Publisher Full Text

41. Heikamp K, Bajorath J: **Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations.** *J Chem Inf Model.* 2013; **53**(4): 791–801.
PubMed Abstract | Publisher Full Text

42. Namasivayam V, Hu Y, Balfer J, *et al.*: **Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns.** *J Chem Inf Model.* 2013; **53**(6): 1272–1281.
PubMed Abstract | Publisher Full Text

43. Gupta-Ostermann D, Hu Y, Bajorath J: **Systematic mining of analog series with related core structures in multi-target activity space.** *J Comput Aided Mol Des.* 2013; **27**(8): 665–674.
PubMed Abstract | Publisher Full Text

44. Dimova D, Stumpfe D, Bajorath J: **Quantifying the fingerprint descriptor dependence of structure-activity relationship information on a large scale.** *J Chem Inf Model.* 2013; **53**(9): 2275–2281.
PubMed Abstract | Publisher Full Text

45. de la Vega de León A, Hu Y, Bajorath J: **Systematic identification of matching molecular series and mapping of screening hits.** *Mol Inf.* 2014; *In press.*

46. Hu Y, Bajorath J: **Many drugs contain unique scaffolds with varying structural relationships to scaffolds of currently available bioactive compounds.** *Eur J Med Chem.* 2014; **76**: 427–434.
Publisher Full Text

47. de la Vega de León A, Bajorath J: **Matched molecular pairs derived by retrosynthetic fragmentation.** *Med Chem Commun.* 2014; **5**(1): 64–67.
Publisher Full Text

48. Hu Y, Bajorath J: **Programs for chemoinformatics and computational medicinal chemistry**. 2014.
Data Source

49. Hu Y, Bajorath J: **Data sets for chemoinformatics and computational medicinal chemistry**. 2014.
Data Source

Open Peer Review

Current Referee Status:

Version 1

Referee Report 22 April 2014

doi:10.5256/f1000research.3979.r4077

Patrick Walters
Vertex Pharmaceuticals Incorporated, Cambridge, MA, USA

The ability to compare multiple computational methods across a series of consistent, high-quality datasets is critical to the progress of computational chemistry and cheminformatics. In the past, each paper published in the field seemed to present yet another new dataset. This dataset heterogeneity made it difficult, if not impossible, to objectively compare methods, and impeded the progress of the field. The availability of large repositories of carefully curated data is critical to the progress of the field. The datasets described in this paper will provide an invaluable resource for future studies. It is refreshing to see the emergence of platforms like ZENODO dedicated to hosting this data.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 17 April 2014

doi:10.5256/f1000research.3979.r4409

Chris J. Swain
Cambridge Med Chem Consulting, Cambridge, UK

Building and testing novel computer models requires access to suitable datasets. The authors have compiled a very useful set of interesting datasets and made them readily available in standard formats (SMILES and SDF). This allows others to both test existing algorithms and to develop new ones.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 13 March 2014

doi:10.5256/f1000research.3979.r4079

Ajay Jain
HDF Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA

Hu and Bajorath offer an update to their resource for computational chemistry. The curated data, and its engineered availability, will be of great interest, especially to methods developers. Even those researchers that are interested in exploring larger data sets that illuminate issues such as activity cliffs and small-molecule structural motifs will find the resource of interest.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Referee Response 17 Apr 2014

Chris J. Swain, Cambridge MedChem Consulting, UK

Such collections of data sets are absolutely invaluable for testing existing algorithms and for developing new ones.

Competing Interests: None