Bioinformatics Advance Access published August 11, 2016

Structural bioinformatics

ChemTreeMap: An Interactive Map of Biochemical Similarity in Molecular Datasets

Jing Lu[1] and Heather A. Carlson[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America, [2]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan, United States of America.

*To whom correspondence should be addressed.

Associate Editor: Prof. Anna Tramontano

Abstract

Motivation: What if you could explain complex chemistry in a simple tree and share that data online with your collaborators? Computational biology often incorporates diverse chemical data to probe a biological question, but the existing tools for chemical data are ill-suited for the very large datasets inherent to bioinformatics. Furthermore, existing visualization methods often require an expert chemist to interpret the patterns. Biologists need an interactive tool for visualizing chemical information in an intuitive, accessible way that facilitates its integration into today's team-based biological research.

Results: ChemTreeMap is an interactive, bioinformatics tool designed to explore chemical space and mine the relationships between chemical structure, molecular properties, and biological activity. ChemTreeMap synergistically combines extended connectivity fingerprints and a neighbor-joining algorithm to produce a hierarchical tree with branch lengths proportional to molecular similarity. Compound properties are shown by leaf color, size, and outline to yield a user-defined visualization of the tree. Two representative analyses are included to demonstrate ChemTreeMap's capabilities and utility: assessing dataset overlap and mining structure-activity relationships (SAR).

Availability: The examples from this paper may be accessed at http://ajing.github.io/ChemTreeMap/. Code for the server and client are available in the Supplementary Information, at the aforementioned github site, and on Docker Hub (https://hub.docker.com) with the nametag ajing/chemtreemap.

Contact: carlsonh@umich.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Researchers in the field of bioinformatics are frequently tasked with exploring the relationship between chemical structures and their potential biological actions. For example, one can predict the interaction between a small molecule and a corresponding protein target, given the chemical structure (Yamanishi *et al.*, 2010). The degree of similarity between chemical structures can also indicate a potential for drug repositioning (Liu *et al.*, 2014; Huang *et al.*, 2015) or off-target effects (Gohlke *et al.*, 2015). A molecules' ability to inhibit protein-protein interactions may also be predicted (Kuenemann *et al.*, 2015). In order to discover such interactions, researchers must harness large databases, including PubChem (Wang *et al.*, 2012) and ChEMBL (Gaulton *et al.*, 2012), which contain vast amounts of heterogeneous biological and chemical data. The size and complexity of these datasets necessitate automated tools to explore available chemical space in order to determine chemical relationships and predict potential interactions.

Exploring chemical space frequently requires a comparison of the number of shared chemicals among multiple databases (Huang *et al.*, 2015), an understanding of the similarity within a compound series (Huang *et al.*, 2015; Gohlke *et al.*, 2015; Liu *et al.*, 2014)series, and an analysis of how compound structures give rise to a specific biological action (Kuenemann *et al.*, 2015). Such analysis begins with the visualization of molecular datasets. There are general visualization strategies such as Venn diagrams (Huang *et al.*, 2015), networks (Huang *et al.*, 2015), heat maps (Gohlke *et al.*, 2015), and clusters (Liu *et al.*, 2014). However, those strategies have limited utility, depending on the biological question and how detailed the analysis must be.

Graphical tools have been developed for cheminformatics that group structurally similar molecules together and display information about molecular structure and bioactivity (Schuffenhauer *et al.*, 2007; Wetzel *et al.*, 2009; Wawer and Bajorath, 2010; Wollenhaupt and Baumann, 2014; Sander *et al.*, 2015). The previous methods usually require users to choose selection rules (Wawer and Bajorath, 2010) and tune parameters (Wollenhaupt and Baumann, 2014). This can limit the utility of these tools for large, diverse sets of data that are the hallmark of bioinformatics. Those tools can also require domain experts to effectively use them, which limits their extension to a larger user base.

To overcome these limitations, we have developed ChemTreeMap, an open-source tool for visualizing compound similarity coupled with associated biochemical information. We have carefully selected techniques for calculating molecular similarity and representing chemical space to satisfy the needs of bioinformaticians. The tool uses standard procedures, requires no tuning parameters, and allows users to interactively explore a molecular dataset. ChemTreeMap organizes molecules into a hierarchical tree based on chemical similarity, much like phylogenetic trees that are commonly used in biology. This provides a familiar framework for scientists to view and access desired information.

Users can map any property of interest to the graph's leaf attributes (i.e. color, size, and border color). This facilitates an on-the-fly, customized exploration of the relationships between molecular structure and other properties. ChemTreeMap's organization reflects the similarities of molecules at various levels and in different chemical series. It does not rely on any assumption about the similarity cutoff. Users can explore the branches to understand the similarity across the nearby molecules. The branch lengths quantify the difference in features, which is particularly useful for a structurally diverse datasets. Longer distances between chemical families highlight more diverse regions of chemical space.

To illustrate ChemTreeMap's utility, we describe two practical applications: the visualization of chemical overlap between molecular datasets and the extraction of structure-activity relationships (SAR).

2 Methods

Organization and visualization of a molecular library requires three considerations: first, how to represent a molecule, second, how to quantify the similarities between different molecules, and lastly, how to represent these similarities graphically. Each ChemTreeMap is then completed by coloring the resulting tree to convey biochemical information.

2.1 Representation of the molecules

Our primary molecular descriptors are stereochemistry-aware, extended connectivity fingerprints (ECFP6#S). These are topological descriptions that capture large, recursive, circular neighborhoods around each atom. This method identifies the functional groups in each molecule, and it is quick to calculate, which makes it well suited for large molecular datasets (Rogers and Hahn, 2010). ChemTreeMap also calculates atom-pairs fingerprints (Carhart *et al.*, 1985) as an alternative metric. Others can be added easily, such as MACCS keys (Durant *et al.*, 2002), topological torsion fingerprints (Nilakantan *et al.*, 1987), and 2D-pharmacophore fingerprints (Gobbi and Poppinger, 1998).

A potential concern about fingerprints is that their pair-wise comparisons may not be optimal for a global description of the data (Wollenhaupt and Baumann, 2014). Fortunately, ChemTreeMap's display uses a global, hierarchical organization to convey information at the local level and in the overall patterns across the data.

2.2 Construction of the chemical similarity tree

The similarity of two molecules is calculated by a Tanimoto coefficient (T_c) (Levandowsky and Winter, 1971), which refers to the number of chemical features they share in common divided by the union of all features (a % similarity with values from 0 to 1). T_c was chosen because of it is fast, easy to implement, and widely used in chem-informatics and drug-discovery software. Branch lengths in ChemTreeMap are inversely proportional to the T_c between the molecules, where shorter branches show high similarity and longer indicate greater diversity.

To build the hierarchical tree, we chose the Neighbor-Joining algorithm (NJ, see Supplementary Information) (Saitou and Nei, 1987). The NJ algorithm is widely used in building phylogenetic trees for large and diverse sequences (Vinh and von Haeseler, 2005; Tamura *et al.*, 2004). It has been mathematically proven that given a correct input distance matrix, the output tree and branch lengths from NJ will also be correct (Mihaescu *et al.*, 2007). Furthermore, NJ does not rely on any parameter tuning, making the tree construction more robust. RapidNJ has a best-case running time of $\boldsymbol{O(N^2)}$ and at worst $\boldsymbol{O(N^3)}$ (Simonsen *et al.*, 2008). It is an agglomerative joining method that follows:

1. $\boldsymbol{D}$ is an $\boldsymbol{N} \times \boldsymbol{N}$ distance matrix, where each element $\boldsymbol{D_{ij}} = T_c(i,j)$ and (i, j represent two molecules from the set of $\boldsymbol{N}$ molecules)
2. The average distance from molecule i to all other molecules k is:
$$\boldsymbol{u_i} = \sum_k \boldsymbol{D_{ik}} / \boldsymbol{N-2}$$
3. With $\boldsymbol{N}$ average distances, we create the $\boldsymbol{Q}$ matrix, with elements:
$$\boldsymbol{Q_{ij}} = \boldsymbol{D_{ij}} - \boldsymbol{u_i} - \boldsymbol{u_j}$$
4. Find the i, j with the smallest value $\boldsymbol{Q_{ij}}$.
5. For that i, j pair, an imaginary ancestor node a is created to replace i and j. The distance between a and i, j is:
$$\boldsymbol{v_i} = \mathbf{0.5} \times \boldsymbol{D_{ij}} + \mathbf{0.5} \times (\boldsymbol{u_i} - \boldsymbol{u_j})$$
$$\boldsymbol{v_j} = \mathbf{0.5} \times \boldsymbol{D_{ij}} + \mathbf{0.5} \times (\boldsymbol{u_j} - \boldsymbol{u_i})$$
6. $\boldsymbol{D}$ is then updated by replacing i and j with an ancestral node a. The distances between a and all the other nodes k are:
$$\boldsymbol{D_{ak}} = (\boldsymbol{D_{ik}} + \boldsymbol{D_{jk}} - \boldsymbol{D_{ij}})/\mathbf{2}$$
7. Keep updating until the last two nodes are joined.

2.3 Clustering of Molecules for Very Large Datasets

For huge datasets, such as ChEMBL (Gaulton *et al.*, 2012), BindingDB (Liu *et al.*, 2007), and ChemBank (Seiler *et al.*, 2008), the molecules need to be clustered by similarity to reduce the size of the task. An interactive display of millions of leafs is not possible with current technology. For tractability, ChEMBL and ChemBank were each independently sorted into their own 4000 clusters. Because BindingDB is smaller, its molecules were clustered into 2000 sets. All clusters were represented by the molecule with the highest sum of T_c across the subset, the molecule most similar to all the others in the group. The size of the leaf for that center is proportional to the number of molecules in the group. It would be possible to produce a ChemTreeMap for each cluster, but wading through thousands of smaller trees is not practical.

MiniBatch-KMeans was chosen for clustering, which has low runtime complexity $\boldsymbol{O(N)}$, memory usage $\boldsymbol{O(N)}$, and relatively low error (Sculley, 2010). The RDKit (*http://www.rdkit.org*, accessed 28 October 2015) fingerprint was selected for this initial clustering, which is an implementation of a Daylight-like fingerprint. The task of finding nearest neighbors is relatively easy, and this alternate fingerprint is fast and saves

memory. With MiniBatch-Kmeans and RDKit fingerprint, the clustering step for ChEMBL – the largest dataset – took less than 5 hours to run on a machine with 16G memory (i3-2100 CPU @ 3.10GHz**).** In comparison, the maximum dissimilarity method (Hassan *et al.*, 1996) implemented in PipelinePilot (Hassan *et al.*, 2006) takes more than a week to cluster ChEMBL, and an algorithm with average runtime complexity $\boldsymbol{O(N \times \log N)}$ such as DBScan still takes more than 3 days.

3 Implementation

The application has a client-server paradigm (Benatallah *et al.*, 2004), in Fig. 1, which allows online access to the results and facilitates sharing data with colleagues. The server performs tree construction, including fingerprint calculations, similarity calculations, neighbor joining, layout optimization, pIC_{50}, ligand efficiency (LE = 1.37×pIC_{50}/heavy-atoms, for rough conversion to ΔG_{bind}/HA), and SlogP calculations (Wildman and Crippen, 1999). All output from the server processes are packaged into a JavaScript Object Notation (JSON) file for client input. The client process displays the tree structures and supports functions, such as searching by ID and changing leaf attributes: circle size, colored border, and biochemical metric mapped to the circle. Layout optimization uses a multi-scale version of dynamic, spring-model layouts implemented in GraphViz (Ellson *et al.*, 2001). This allows users to actively pull and reorient nodes/branches to improve the visual layout in a local area.

ChemTreeMap is a web-based tool, which is easy to set up on any computer with Python 2.7. GraphViz and RapidNJ are freely available online for major operating systems (Ellson *et al.*, 2001; Simonsen *et al.*, 2008). The graphics can be viewed on any computer with an HTML5 capable browser (tested on Google Chrome Version 46, Internet Explorer 11, and Safari 9.0). We recommend Google Chrome for displaying ChemTreeMaps of thousands of molecules because of the computational intensity of its force-directed graph. Speed also depends on client hardware.

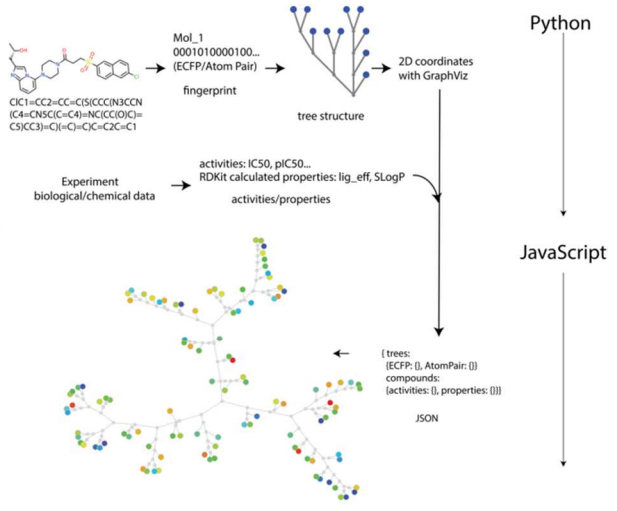

Fig. 1. ChemTreeMap work flow. Each compound (encoded as a standard SMILE string) and its biological data are processed in steps to yield a JSON file, which is then used by a JavaScript App to create the graphic tree map (clearly three independent chemical series in this sample). Specifically, ECFP6#S and Atom-Pair fingerprints are calculated as two options for the T_c used in building a tree structure. By default, LE and SlogP are calculated for each molecule and its activity data. Tree structures and compound data with associated properties are required in the JSON file.

3.1 ChemTreeMap can be easily extended with more features

A straightforward option to add functionality is to input additional data for each molecule, as columns in a tab-delineated input file (e.g. SMILE string, data 1, data 2, etc.). The program will present the additional data as options in drop-down menus for the user to map that information onto the ChemTreeMap leafs through display options (color, border, size).

A more powerful alternative is to extend the functionality in ChemTreeMap's TreeBuild.py file. Currently, the chemical properties of LE and SlogP are calculated using RDKit in TreeBuild.py, and developers can very easily add 55 additional descriptors from RDKit. For more descriptors, we recommend adding data from MOE (Chemical Computing Group Inc., 2014) or PaDEL (Yap, 2011) with the simple tab-delimited option. Any in-house, custom analysis is best added directly in the python code if it will be used frequently.

3.2 Comparing ChemTreeMap to other SAR methods

We compared ChemTreeMap with four freely available programs: Similarity-Potency Tree (SPT, Wawer and Bajorath, 2010), Data Warrior (Sander *et al.*, 2015), Scaffold Hunter (Wetzel *et al.*, 2009), and CheS-Mapper (Gütlein *et al.*, 2012, 2014). SPT is the method most similar to ChemTreeMap. It uses T_c with ECFP4#S fingerprints to calculate chemical similarity, but it uses a different type of tree display and network analysis. SPT has a compound-centric view, focusing on potency and a limited set of nearest neighbors in chemical space. Their tree structures rely on a chosen reference compound. Similar molecules are not necessarily grouped together in some cases.

Data Warrior is an open-source program with highly interactive graphical views and interesting statistical analysis. It focuses on similar neighborhood relations only and ignores similarity relations below a certain threshold. Such methods may not capture enough structural breadth for a diversified dataset.

Scaffold Hunter identifies chemical cores for each molecule and associates them in a hierarchy based on medicinal chemistry rules. A tree-like output is used, but the local SAR of individual molecules sharing the same scaffold is not displayed.

CheS-Mapper focuses on analysis of QSAR models and extendibility to other applications is not clear. It uses chemical clustering and dimensionality reduction to organize compounds into a virtual 3D space. The space is defined by physicochemical descriptors of chemical/biological similarity that are chosen by the user. Unfortunately, the compression of multiple descriptors into three dimensions inherently produces a loss of information (the same likely applies to ChemTreeMap). More importantly, many of the descriptors are not well known in the bioinformatics community. This tool is best for a chemist who knows how to choose the best options. Fortunately, statistical analysis is provided to interpret the degree of lost data, so the non-expert is at least warned of poor choices.

We did not have access to a maximum common substructure (MCS) method for comparison. MCS is another way to organize molecules into groups, based on pre-defined chemical functionalities and novel pattern matching (Gardiner *et al.*, 2007). Like Scaffold Hunter, MCS techniques identify common core patterns across a dataset. InSARa is a recently introduced MCS method (Wollenhaupt and Baumann, 2014) that uses reduced graphs to create tree-based output like ChemTreeMap, SPT, and Scaffold Hunter. Its final representative substructures are sensitive to many parameter choices (Wollenhaupt and Baumann, 2014), which is where ChemTreeMap and inSARa differ. Our generalized T_c-based distances have no tunable parameters, other than the choice of fingerprints. A tunable method has its own benefits, and we see inSARa as a complementary method to ChemTreeMap, SPT, and Scaffold Hunter.

4 Datasets

ChemTreeMap is applicable to a wide range of datasets with various levels of complexity and sizes. To demonstrate ChemTreeMap, we have chosen diverse biomolecular datasets ranging from thousands to millions of molecules. To show chem-data overlap, we use some of the largest datasets with bioactivity data: ChEMBL v. 20 (Gaulton *et al.*, 2012), BindingDB (Liu *et al.*, 2007), and ChemBank (Seiler *et al.*, 2008).

For SAR examples, we assembled chemical datasets for the four protein targets shown in Table 1. Clotting factor Xa (FXa), cyclin-dependant kinase 2 (CDK2), p38α MAP kinase (p38α), and cytochrome P450 3A4 (CYP3A4) were chosen because their SAR are well characterized (Sutherland *et al.*, 2004; Fontaine *et al.*, 2005), and they have been used in previous studies for visualizing chemical data (Wawer and Bajorath, 2010; Wollenhaupt and Baumann, 2014). The data for FXa, CDK2, and p38α are from BindingDB. CYP3A4 data (bioassay AID:884) was pulled from high-throughput screening (HTS) data in PubChem (Wang *et al.*, 2012). All four systems have a large range of inhibition data. The average T_c for each set is under 0.2, which indicates high chemical diversity.

Each of the SAR datasets was prepared using the following protocol:

1. For inclusion, a molecule must have an IC_{50} or pIC_{50} for bioactivity. For PubChem data, only molecules in the "Active" category were kept for analysis.
2. If there were multiple activity data for a molecule, the average of the IC_{50} was used (i.e. no repeat chemical structures).
3. The ionization state and tautomers for each chemical structure were determined using the "wash" utility in MOE 2014 (Chemical Computing Group Inc., 2014).

Table 1. Datasets used for SAR analysis.

Protein Target	Target class	Number of Molecules	Max pIC_{50}	Min pIC_{50}	Ave T_c
FXa	protease	2161	10.7	3.0	0.172
CDK2	kinase	1923	9.5	2.9	0.141
p38α	kinase	5139	10.4	2.9	0.167
CYP3A4	oxidoreductase	6837	8.9	4.1	0.115

5 Results and Discussion

5.1 Comparing chemical diversity of large datasets

When more data is needed for a project, information that covers new chemical space is typically preferred. Depending on the project, very wide diversity may be desirable to enhance breadth, or new compounds in nearby chemical space may be needed to increase the depth of coverage. ChemTreeMap can identify both types of chemical similarity/diversity.

Existing methods count the number of duplicate molecules between the sets (Huang *et al.*, 2015), or they map chemical features to a 2D scatter plot with overlapping regions, based on principal component analysis or multidimensional scaling (Awale *et al.*, 2013; Lewis *et al.*, 2015). These methods can fail to convey enough information on the structural similarity within/between the molecular sets.

Our examples for comparing large sets revolve around ChEMBL (Gaulton *et al.*, 2012). It is a dataset with ~1.3 million chemicals with significant biochemical annotation from the literature (~50 journals). We compared ChEMBL to ChemBank and BindingDB to show datasets that represent collections with more breadth vs more depth, respectively. ChemBank (Seiler *et al.*, 2008) is a set of ~1.15 million molecules constructed from chemical HTS studies. Many compounds in ChEMBL are not from HTS studies and will not appear in ChemBank. Though many HTS studies eventually appear in the literature, the full data for every compound rarely appears in a final publication. Therefore, we knew that many molecules would appear in one but not the other. We did not know how similar/different the chemical sets were from one another, but we expected significant populations of ChEMBL-only and ChemBank-only compounds with many in similar chemical space.

The traditional Venn diagram in Fig. 2 shows that the overlap is relatively small as expected; 11.3% of ChEMBL molecules are the same as 12.8% of ChemBank's compounds. To show the chemical similarity and diversity, a more detailed display is needed. ChemTreeMap highlights regions where the chemical space is unique to ChemBank (red branches A and B) and to ChEMBL (blue branch C). Though some branches contain red and blue leafs in similar chemical space, combining the two sets primarily increases the breath of chemistry structures overall.

ChEMBL (Gaulton *et al.*, 2012) and BindingDB (Liu *et al.*, 2007) both curate molecules from the biochemical literature. BindingDB focuses on protein targets with crystal structures in the Protein Data Bank. It enhances the structures with experimental binding data for many ligands. ChEMBL is a major source of data for BindingDB's molecules, but BindingDB includes data from 12 biochemical journals not covered by ChEMBL. Together, ChEMBL and BindingDB's shared efforts provide ~60 journals-worth of data to the scientific community.

Based on their construction, a large overlap is expected for ChEMBL +BindingDB (Fig. 3). About 82% of the molecules in BindingDB are also in ChEMBL (black nodes). No large branches are dominated by BindingDB. Instead, many of ChEMBL's branches contain new molecules from BindingDB. Fig. 3 shows that this augmentation to ChEMBL from BindingDB adds depth of coverage, specific to drug-like space.

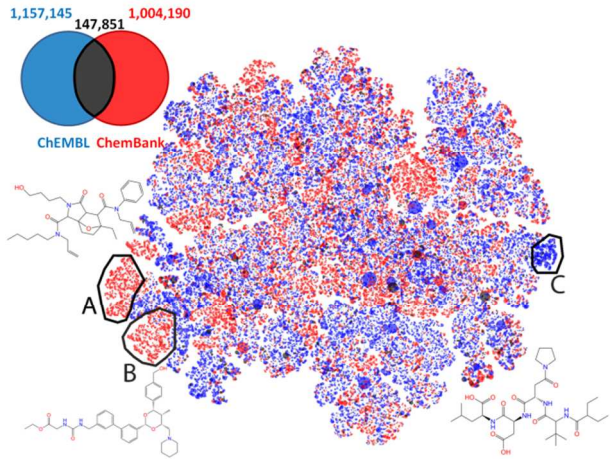

Fig. 2. Breadth. The addition of ChEMBL (blue) and ChemBank (red) adds breadth of chemical space. ChemTreeMap details the regions of chemical similarity and diversity for both sets. Shared molecules are in black clusters. Regions A and B contain molecules only from ChemBank. Region C contains molecules only from ChEMBL. Representative molecules are shown. Leaf size (circles) is proportional to the number of molecules in each cluster from the initial processing step for large databases. The view is scaled out to show the entire data space. Full details can be seen by zooming in on any region.

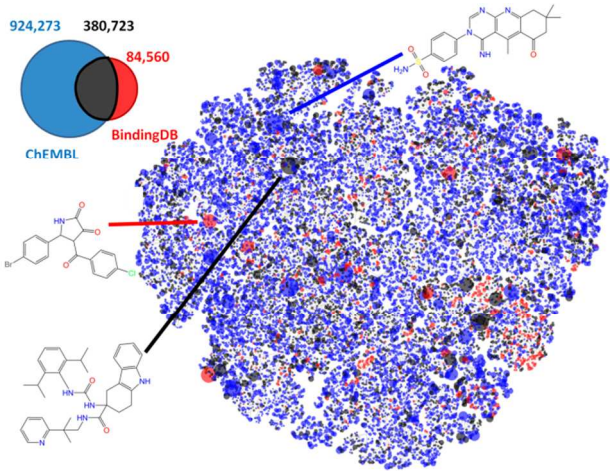

Fig. 3. Depth. The addition of ChEMBL (blue) and BindingDB (red) adds depth to chemical space: Note that the ChEMBL-only branches are for targets that do not have a protein-ligand crystal structure, a requirement for BindingDB. Shared molecules are in black. Representative molecules are shown. The view is zoomed out to show all data.

5.2 Options for displaying more information

A distinct advantage for ChemTreeMap is the ability to display multiple layers of information simultaneously through leaf color, outline, and size. For example, molecules with good solubility and high LE are desirable for drug leads. Fig. 4 shows how ChemTreeMap adds more data through outlines. Molecules with good solubility (low number, red) or good LE (higher number, red) can be easily recognized in the graph. Users could extend ChemTreeMap with more SAR metrics, like SALI scores (Structure-Activity Landscape Index, Guha and Van Drie, 2008) or PAINS alerts (Pan Assay Interference compounds, Baell and Holloway, 2010).

ChemTreeMap uses a broad color scheme purple-blue-cyan-green-yellow-red. Previous studies have scaled the colors green-yellow-red to fit the max/min range in each dataset (Wawer and Bajorath, 2010; Wollenhaupt and Baumann, 2014). Having twice the number of colors allows us to show a larger span of properties with better clarity. It also allows users to easily compare different datasets because the static colors always indicate the same IC_{50} in any ChemTreeMap. The spectrum basically covers the full range of affinities obtained in biological assays, where inhibitors vary in IC_{50} from ~100 µM (purple: $pIC_{50} \geq 4$) to ~1 nM (red: $pIC_{50} \leq 9$). For LE, the range is 0 (purple) to 0.5 (red) kcal/mol-HeavyAtom, which covers 90% enzyme activity data from our previous study (Carlson *et al.*, 2008). For SlogP, the range is set from -5 (red) to 5 (purple). Users can switch between attributes while inspecting the branches, and the features are interactive, e.g. navigating, dragging, or zooming to see details.

5.3 ChemTreeMap can be used to extract SAR information

For the rest of our examples, the molecules were not clustered, and all leafs represent a single compound. Traditionally, SAR information is deduced by analyzing molecules from several chemical series, often incorporating predictive statistical models (Wawer and Bajorath, 2010). Interpretation can be strongly dependent on the experience of the chemists, and these paradigms are limited in the number of compounds one can analyze. ChemTreeMap provides an intuitive, robust, and effective way to extract SAR information from a chemical library. ChemTreeMap does not rely on any assumption about the similarity cutoff. Its hierarchical organization of molecules, based on structural similarity, facilitates the identification of SAR hotspots and activity cliffs. The distance between leafs highlights the (dis)similarity between any molecule pair.

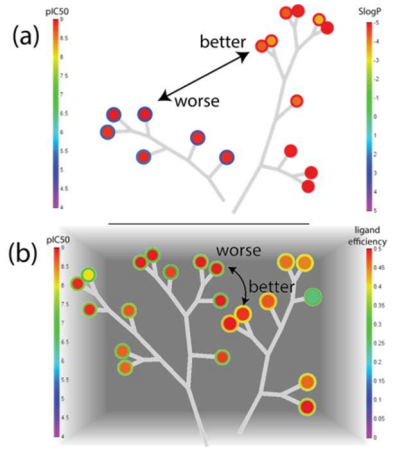

Fig. 4. Multi-layer information. (a) pIC_{50} color inside the circles and SlogP mapped to the outline color. The left branch shows a common pitfall of good affinity but poor solubility of the molecules. (b) Gray background added to make outlines more visible. pIC_{50} color inside the circles and LE mapped to outline color. Most leafs have high potency, but those with green outlines are larger molecules with lower LE. These molecules are less "efficient" because they have the same binding affinity but need more contacts to the target.

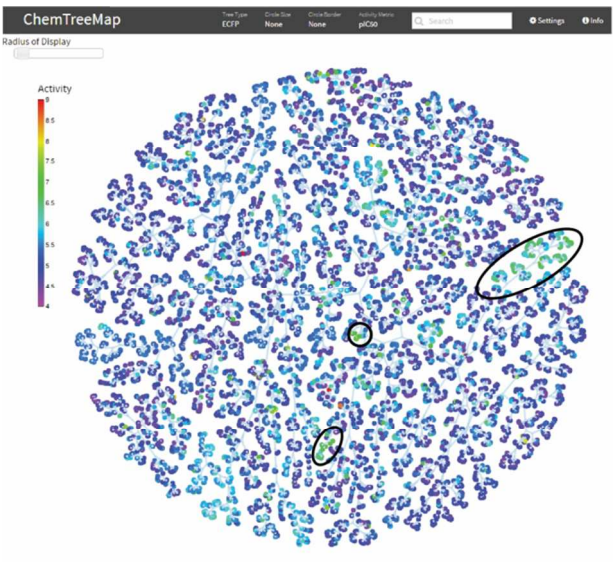

Fig. 5. ChemTreeMap for CYP3A4. The set (6837 molecules) shows a few compact regions of chemical space with moderate activity (green leafs, highlighted by black ovals). Compounds in these groups have a higher potential for drug development. There are also a small number of one-off, strong inhibitors that are likely HTS error (rare orange and red leafs). The top bar contains drop down menus for ECFP or Atom-Pair fingerprints and the leafs (circle size, circle border, activity metric, settings, and info). The tree is dynamic using a force-directed graph with control of the "Radius of Display" (see Supplementary Information). The color bar shows the activity metric. Each color represents one level of potency (pIC_{50}): which is 4 (purple, IC_{50} = 100 µM, indigo), 5 (blue), 6 (cyan, IC_{50} = 1 µM,), 7 (green), 8 (yellow), 9 (IC_{50} = 1 nM, red).

5.3.1 HTS

HTS is a screening technology that tests thousands of molecules in a biochemical assay. Hits are typically inhibitors of an enzymatic assay, but many different assays exist that test a variety of effects. ChemTreeMap displays can help with interpreting the data. ChemTreeMap has an advantage when exploring large, heterogeneous datasets from HTS because of its speed, hierarchical structure, and broad color range.

Most of the compounds used in HTS are inactive, which can be seen by the predominance of blue/purple in a CYP3A4 ChemTreeMap, Fig. 5. HTS data is notorious for many false positives and false negatives. One of the hallmarks of true positives is finding many similar molecules displaying moderate activity like those marked in Fig. 5. These regions indicate chemical space with potential for further development. Groupings of hits can be a small, 5-molecule sub-tree or a large branch of 30+ compounds. Of course, proper statistics are critical to assessing signal-to-noise in HTS data. One way that ChemTreeMap could be easily extended for a custom HTS application is to map statistical significance of each compound's signal onto their leaf outlines (e.g. z-score).

5.3.2 SAR of FXa

Fig. 6 is the ChemTreeMap of our SAR set for FXa. A typical SAR set explores a more focused area of chemical space (inhibitors in this example). The molecules have a range of activities that change with the structural features. Branches dominated by strong inhibitors in red clearly denote specific chemistry linked to high potency.

Sub-tree I is the largest region of the ChemTreeMap with high activity. The neighboring, dashed region shows how potency starts to decrease as larger chemical modifications are made. Fig. 7(a) shows a magnified view of sub-tree I from Fig. 6, with LE data shown on the outlines of the leafs. By inspecting the neighboring molecules, we can identify shared chemical features that are correlated with high activity. Fig. 7(b) shows that all molecules A-L in sub-tree I share a 2-(4-(N,N-dimethyl carbamimidoyl)benzamido)-N-(pyridin-2-yl)benzamide core, marked in gray. *ChemTreeMap makes a rather complex chemical analysis straightforward for informaticians because the pattern is clear from the visual display.* An "activity cliff" is easy to identify for compounds J and K. Activity cliffs occur when a small chemical modification leads to a large change in activity (Bajorath, 2014). ChemTreeMap makes it easy to find molecules in very close proximity with large color changes. Several other structural features are also seen in the full dataset, such as activity switches and SAR hotspots also identified using inSARa (Wollenhaupt and Baumann, 2014), but for brevity, we focus on region I.

Molecules A-H/J/K differ by small chemical changes on the central ring. Molecule I differs from H by a methyl group in place of a chlorine at the bottom of the molecule. I is a "descendent" of H because they are most similar, but I is one step away from the common structure of the other compounds. The addition of a four-membered ring in L is a large chemical change that places it in a nearby sub-tree separate from A-K.

Fig. 8 provides the analyses of the FXa dataset with SPT, Data Warrior, and Scaffold Hunter. Here, we focus the discussion on compounds from sub-tree I. SPT is the state-of-the-art method for cheminformatics. An accurate method should show agreement with SPT and improvement where possible. Fig. 8(a) shows that the same molecules from sub-tree I gave a large structural feature in the highest-ranked tree found with SPT, which is in good agreement with our finding that sub-tree I is the largest

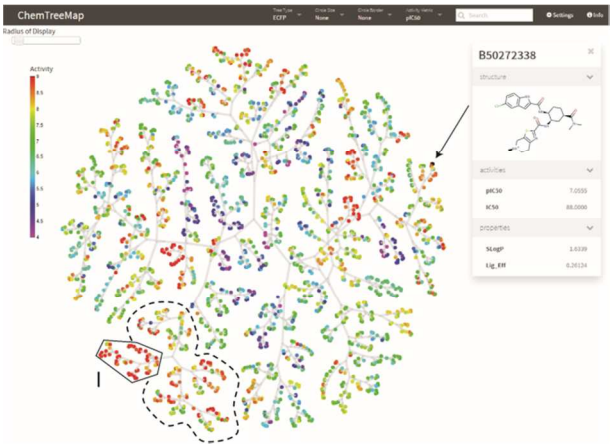

Fig. 6. ChemTreeMap for 2161 FXa inhibitors: Sub-tree I contains high potency molecules; the dashed region highlights many modifications in the nearby chemical space cause a drop in potency. Inspection of those compounds shows that the most detrimental changes involve removing a positive charge from an essential functional group. The color bar follows the same activity pattern as in Fig. 5. The upper-left region shows that clicking on a leaf provides the chemical information for that molecule.

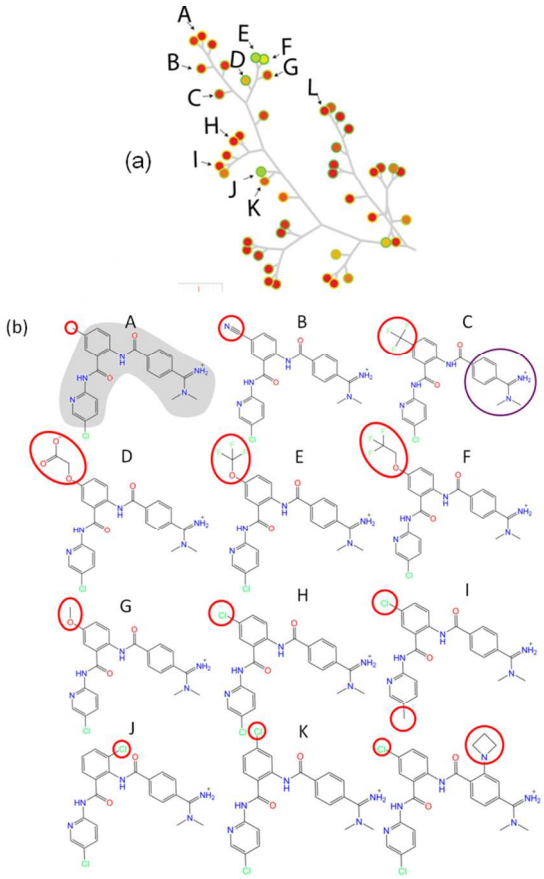

Fig. 7. Detail for the FXa map. (a) Sub-tree I from Fig. 6 with activities as fill color and LE as outline color. Inhibitors with larger functional groups have less favorable LE values (compounds D-F). (b) The chemical structures of representative molecules A-L are shown. The common core is marked in gray, and the differences in the functional groups are noted with red circles. The functional group circled in purple is an essential feature for these FXa inhibitors, and L is a representative of a nearby sub-tree that modifies this part of the molecule.

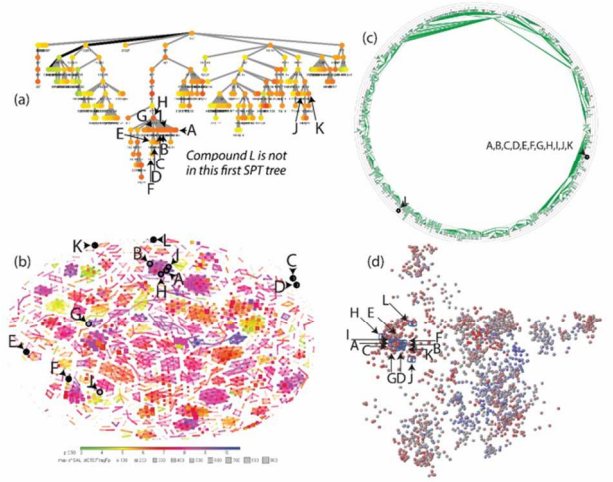

Fig. 8. FXa with other tools. Sub-tree I of the FXa dataset visualized using four other tools: (a) SPT, (b) Data Warrior, (c) Scaffold Hunter, and (d) CheS-Mapper.

contiguous region of high activity. The majority of compounds in Fig. 8(a) are from the two marked regions in Fig. 6. SPT constructs a different visual tree that uses the compounds as nodes. All molecules with T_c >0.4 of the root compound are organized into levels. Each level shows all molecules with T_c >0.55 to the root-molecule above it. SPT organizes each level by increasing activity from left to right, but removes the structural relationships between molecules in the same level. The ordering directs the user toward the most active compounds, but it may obscure structural features of the SAR. For instance, H/J/K only differ by the location of one chlorine atom, yet they appear unrelated in Fig. 8(a).

In Fig. 8(b), Data Warrior spreads A-L across its entire network, with few connections between them. Several of the molecules are shown as lone data points. In Fig. 8(c), Scaffold Hunter correctly identifies the common substructure of A-K, but molecule L is placed in another tree with cyclobutane as a root. The relationship between A-K and L is lost. The impact of small modifications on activity is not presented, but the information is necessary for SAR studies. In Fig. 8(d), CheS-Mapper gives relationships similar to ours. D-G are close together and H-L have the same pattern, but C is grouped with A+B in CheS-Mapper not with D-G like ChemTreeMap. Either combination is reasonable. For the full dataset, CheS-Mapper was overcrowded, and some SAR was obscured.

Clearly, future enhancements for ChemTreeMap should include scaffold information like Scaffold Hunter. The ancestor nodes that serve as branch points are prime locations for showing the shared common substructure of the descendent molecules. It would also be an appropriate point in the graph to display SAR alerts based on the patterns of activity in the descendent branches. Though not seen in the data for Fig 8(d), CheS-Mapper has alerts for activity cliffs, and similar alerts should be incorporated into ChemTreeMap.

5.3.3 SAR hotspot in CDK2 data

An SAR hotspot is a collection of similar molecules displaying a wide range of potency. Finding these series are important for SAR, but they can also facilitate "patent busting." ChemTreeMap makes finding SAR hotspots easier by having 1) a wide color range for activity and 2) branch lengths between nodes that are proportional to chemical similarity. Short branches with many colors are candidates for SAR hotspots.

In Fig. 9(a), the ChemTreeMap for CDK2 contains sub-tree II, an SAR hotspot. In Fig. 9(b), two sets of molecules can be seen in different branches because of their chemical similarity: A/B (ether substitutions) and C-E (alkane substitutions). These are separate from F-H which have larger chemical differences between each other and the rest of sub-tree II. If these molecules were ordered by potency, these groupings would be more difficult to identify. However, the preference for a branched, 4-atom substituent would still be clear.

In Fig 10(a), SPT identifies sub-tree II compounds in its second-ranked tree. The rank of the tree is fitting because some compounds have low activity. The chemical similarity is clear in the SPT tree, but SPT's color scheme is a limitation. The colors are scaled to span the min to max values across the whole dataset. ChemTreeMap's color scale makes it easy to identify the drop in activity for compounds B and F. Data warrior in Fig 10(b) captures the similarity of the ethers in A and B, but they are far from the rest of the compounds. Also, C-E are not necessarily distinct from F-H. Scaffold Hunter in Fig 10(c) shows that all molecules of sub-tree II share the same scaffold, but the SAR between them is not shown. Fig 10(d) shows CheS-Mapper is similar to ChemTreeMap, except C is more "central" and H is too far from A-G. Better detail can be seen in the Supplementary Information.

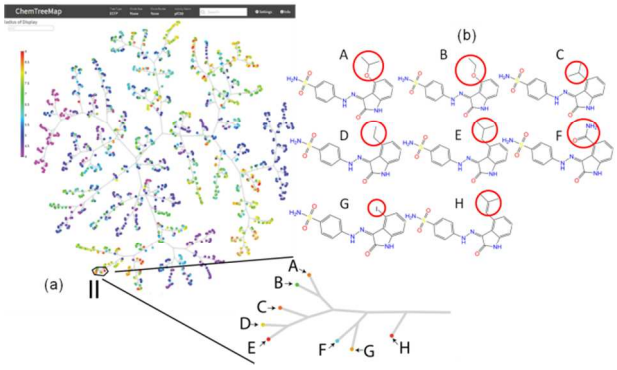

Fig. 9. CDK2. (a) ChemTreeMap for the CDK2 dataset. Sub-tree II contains molecules with large variance in bioactivity. (b) Expanded view of sub-tree II. Red circles mark the differences in the chemistry across the similar molecular cores.

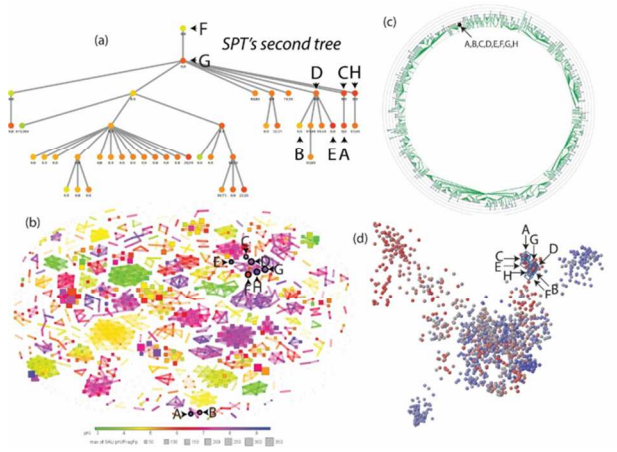

Fig. 10. CDK2 with other tools. The visualization of CDK2 compounds using four other tools: (a) SPT, (b) Data Warrior, (c) Scaffold Hunter, and (d) CheS-Mapper.

Our analysis is based on global exploration of each dataset, but above, our discussions focus on local sub-trees. The Supplementary Information contains diagrams for each method, using only the sub-tree compounds. Minor reorganizations are expected, but the results stay the same overall.

Results for p38α are also in the Supplementary Information.

Conclusion

ChemTreeMap is innovative for quantifying chemical similarity in the branch lengths as done for phylogenetic trees, organizing molecules in an alternative hierarchy, and mapping multiple properties to graphical attributes. It uses robust, widely accepted methods.

ChemTreeMap is designed as a general purpose chemical visualization and data mining tool with many interactive features to ease the navigation in large datasets (dragging, zooming, searching, etc). A single click on a leaf yields detailed molecular information. Its dynamic layout allows users to modify the tree, using a click-and-drag feature on the nodes that can reposition branches to improve the view.

ChemTreeMap is applicable to a wide range of problems. Any data can be mapped onto the similarity tree. The approach does not make any assumptions about the relationship between activity and structure, thus enabling a data-driven interpretation of biochemical information. It can also be also implemented in a client-server format that allows efficient data sharing between collaborators.

Acknowledgements

We thank Chemical Computing Group for their generous donation of the MOE software package. JL thanks the entire Carlson lab for constructive feedback.

Funding

This work was supported in part by the National Institutes of Health (U01 GM086873).

Conflict of Interest: none declared.

References

Awale,M. and Reymond,J.L. (2013) MQN-Mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.*, **53**, 509–518.

Baell,J.B. and Holloway,G.A. (2010) New substructure filters for removal of pan assay interference compounds [PAINS] from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.

Bajorath,J. (2014) Exploring activity cliffs from a chemoinformatics perspective. *Mol. Inform.*, **33**, 438–442.

Benatallah,B. *et al.* (2004) Web service conversation modeling: a cornerstone for e-business automation. *IEEE Internet Comput.*, **8**, 46–54.

Carhart,R.E. *et al.* (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **25**, 64–73.

Carlson,H.A. *et al.* (2008) Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J. Med. Chem.*, **51**, 6432–6441.

Chemical Computing Group Inc. (2014) Molecular Operating Environment (MOE), 2014.10., Montreal, Canada.

Durant,J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.

Ellson,J. *et al.* (2001) Graphviz — open source graph drawing tools. In, *Lecture Notes in Computer Science*. Springer-Verlag, pp. 483–484.

Fontaine,F. *et al.* (2005) Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *J. Med. Chem.*, **48**, 2687–2694.

Gardiner,E.J. *et al.* (2007) Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *J. Chem. Inf. Model.*, **47**, 354–366.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Gobbi,A. and Poppinger,D. (1998) Genetic optimization of combinatorial libraries. *Biotechnol. Bioeng.*, **61**, 47–54.

Gohlke,B.O. *et al.* (2015) 2D and 3D similarity landscape analysis identifies PARP as a novel off-target for the drug Vatalanib. *BMC Bioinformatics*, **16**, 308.

Gütlein,M. *et al.* (2014) CheS-Mapper 2.0 for visual validation of (Q)SAR models. *J. Cheminformatics*, **6**, 1–18.

Gütlein,M. *et al.* (2012) CheS-Mapper - chemical space mapping and visualization in 3D. *J. Cheminformatics*, **4**, 1–16.

Guha,R. and Van Drie,J.H. (2008) Structure--activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.*, **48**, 646–658.

Hassan,M. *et al.* (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol. Divers.*, **10**, 283–299.

Hassan,M. *et al.* (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Divers.*, **2**, 64–74.

Huang,H. *et al.* (2015) DMAP: a connectivity map database to enable identification of novel drug repositioning candidates. *BMC Bioinformatics*, **16**, S4.

Kuenemann,M.A. *et al.* (2015) An exploration of the 3D chemical space has highlighted a specific shape profile for the compounds intended to inhibit protein-protein interactions. *BMC Bioinformatics*, **16**, A5.

Levandowsky,M. and Winter,D. (1971) Distance between sets. *Nature*, **234**, 34–35.

Lewis,R. *et al.* (2015) Synergy maps: exploring compound combinations using network-based visualization. *J. Cheminformatics*, **7**, 36.

Liu,R. *et al.* (2014) Exploiting large-scale drug-protein interaction information for computational drug repurposing. *BMC Bioinformatics*, **15**, 210.

Liu,T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Mihaescu,R. *et al.* (2007) Why neighbor-joining works. *Algorithmica*, **54**, 1–24.

Nilakantan,R. *et al.* (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.*, **27**, 82–85.

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sander,T. *et al.* (2015) DataWarrior: an open-source program For chemistry aware data visualization and analysis. *J. Chem. Inf. Model.*, **55**, 460–473.

Schuffenhauer,A. *et al.* (2007) The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.*, **47**, 47–58.

Sculley,D. (2010) Web-scale K-means clustering. In, *Proceedings of the 19th International Conference on World Wide Web*, WWW '10. ACM, New York, NY, USA, pp. 1177–1178.

Seiler,K.P. *et al.* (2008) ChemBank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.

Simonsen,M. *et al.* (2008) Rapid neighbour-joining. In, Crandall,K.A. and Lagergren,J. (eds), *Algorithms in Bioinformatics*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 113–122.

Sutherland,J.J. *et al.* (2004) A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.*, **47**, 5541–5554.

Tamura,K. *et al.* (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 11030–11035.

Vinh,L.S. and von Haeseler,A. (2005) Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinformatics*, **6**, 92.

Wang,Y. *et al.* (2012) PubChem's bioassay database. *Nucleic Acids Res.*, **40**, D400–D412.

Wawer,M. and Bajorath,J. (2010) Similarity−potency trees: a method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.*, **50**, 1395–1409.

Wetzel,S. *et al.* (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.*, **5**, 581–583.

Wildman,S.A. and Crippen,G.M. (1999) Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.*, **39**, 868–873.

Wollenhaupt,S. and Baumann,K. (2014) inSARa: intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. *J. Chem. Inf. Model.***54**, 1578-1595.

Yamanishi,Y. *et al.* (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**, i246–i254.

Yap,C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.