Journal of Cheminformatics

SOFTWARE Open Access

Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods

Sereina Riniker[†] and Gregory A Landrum[*†]

Abstract

Fingerprint similarity is a common method for comparing chemical structures. Similarity is an appealing approach because, with many fingerprint types, it provides intuitive results: a chemist looking at two molecules can understand why they have been determined to be similar. This transparency is partially lost with the fuzzier similarity methods that are often used for scaffold hopping and tends to vanish completely when molecular fingerprints are used as inputs to machine-learning (ML) models. Here we present similarity maps, a straightforward and general strategy to visualize the atomic contributions to the similarity between two molecules or the predicted probability of a ML model. We show the application of similarity maps to a set of dopamine D3 receptor ligands using atom-pair and circular fingerprints as well as two popular ML methods: random forests and naïve Bayes. An open-source implementation of the method is provided.

Keywords: Visualization, Machine-learning, Similarity, Fingerprints

Background

Chemical structures are often represented by molecular fingerprints where structural features are converted to either bits in a bit vector or counts in a count vector. This abstract representation allows the computationally efficient handling and comparison of chemical structures. Using such fingerprints, the similarity between two molecules can be calculated in a straightforward manner with simple similarity metrics such as Tanimoto [1], Dice [2], and so on. However, depending on the descriptors used to generate the fingerprints, the interpretation of the resulting similarity may not be trivial. This problem worsens when machine-learning (ML) models are trained to predict the activity (or other properties) of new compounds: ML models often appear as complete "black boxes" that just output numeric predictions to their users. Though these predictions can be quite accurate, it has been shown that supplementing numeric predictions with additional information from the model can improve the ability of both expert and non-expert users to work with the results [3]. This provides substantial motivation for the development of strategies to visualize the parts of a molecule contributing to a similarity value or model prediction.

Few visualization approaches for such models are described in the literature. An early example is the visualization of a modal fingerprint [4,5], which contains all bits which are present in 50 - 100% of the molecules of a training set. The atoms are colored based on the similarity to the modal fingerprint, i.e. how many of the bits set by the atom are present in the modal fingerprint. Franke *et al.* [6] visualized the importance of three-point pharmacophores (3PP) obtained from a trained support vector machine (SVM) model by placing differently sized spheres at the centre of the substructure leading to a 3PP. The importance of each 3PP was calculated based on the difference of SVM prediction for a molecule when this 3PP is removed. The interpretation of linear SVM models was also the goal of the heat map coloring scheme developed by Rosenbaum *et al.* [7]. The SVM model was trained using ECFP fingerprints and the authors focussed solely on the coloring of bonds. The coloring was based on the weights obtained from the SVM model, where the final weight of a bond is the normalized sum of

*Correspondence: gregory.landrum@novartis.com
[†]Equal Contributors
Novartis Institutes for BioMedical Research, Basel, Switzerland

the weights of the fingerprints features containing this bond. The color scheme was chosen such that red corresponds to the negative class and green to the positive class with orange as zero. Another approach is the Glowing Molecule visualization which has been used to show the regions of a molecule which may have the most influence on ADME and physicochemical properties [8,9]. A red glow indicates that this region has a positive influence on the property (i.e. the property value increases) while a blue glow indicates a negative influence with green representing no significant overall effect. Unfortunately, a detailed description of the algorithm used for the Glowing Molecule method were not provided and, since it is implemented as part of a commercial product, the method is not generally available.

Here, we present similarity maps, a general approach for the visualization of both fingerprint similarities between two molecules and machine-learning (ML) model predictions. In our scheme, the "weight" of an atom is the similarity or predicted-probability difference obtained when the bits in the fingerprint corresponding to the atom are removed, similar to the approach of Franke *et al.* [6]. The normalized weights are then used to color the atoms in a topography-like map with green indicating a positive difference (i.e. the similarity or probability decreases when the bits are removed) and pink indicating a negative difference, gray represents no change. The visualization is demonstrated for atom pairs and several types of circular fingerprints and subsequently used to explain the factors leading to the predicted probability of a random forest and a naïve Bayes model. All source code and data required to reproduce the examples is provided in the Additional file 1.

Implementation

A "weight" is determined for each atom of the test molecule by removing the bits which are set by the atom in the fingerprint of the test molecule, recalculating the similarity between the modified fingerprint and the fingerprint of the reference compound s_{mod}, and calculating the difference to the original similarity $\Delta s = s_{orig} - s_{mod}$. The fingerprints are calculated using the open-source cheminformatics toolkit RDKit [10]. Dice [2] similarity is used in the current implementation but any other similarity metric could be employed. For AP (a count vector), the bits of an atom i are straightforward to determine, the count for each pair involving atom i is decreased by one. In circular fingerprints, on the other hand, bits are set for different atomic environments, starting at radius 0 up to the maximum radius. In RDKit, the environment (i.e. centre atom and radius) associated with each bit in a fingerprint can be obtained when generating the fingerprint. This information is used to determine all the bits where the atom is part of the environment.

The procedure to calculate "atomic weights" for the similarity between two molecules *ref_mol* and *this_mol* is shown in pseudocode below,

```
ref_fp = calculate_fingerprint(ref_mol)
this_fp = calculate_fingerprint(this_mol)
weights = []
orig_simil = dice_similarity(ref_fp, this_fp)
for atom in this_mol.get_atoms():
   new_fp = calculate_fingerprint_without_atom
            (this_mol, atom)
   new_simil = dice_similarity(ref_fp, new_fp)
   weight = orig_simil - new_simil
   weights.append(weight)
```

Similarity maps can also be used to visualize the atomic contributions to the predicted probability of a ML model. The generation of the bitmap is the same as before, depending on the kind of basic fingerprint used to train the ML model. However, the "atomic weights" are no longer similarity differences but predicted-probability differences,

```
this_fp = calculate_fingerprint(this_mol)
weights = []
orig_proba = predict_model_probability(this_fp)
for atom in this_mol.get_atoms():
   new_fp = calculate_fingerprint_without_atom
            (this_mol, atom)
   new_proba = predict_model_probability(new_fp)
   weight = orig_proba - new_proba
   weights.append(weight)
```

In the case of NB, the difference between the logarithmic probabilities is used. The ML methods were calculated using the open-source toolkit scikit-learn [11].

To construct a similarity map, the atom weights are normalized by dividing by the maximum absolute weight value and then used to calculate bivariate Gaussian distributions centered at the corresponding atom positions. The atom weights influence only the peak and not the variance of the Gaussian distribution. The RDKit function for this makes use of the Python library matplotlib [12]. The similarity map is then generated by superimposing the atom coordinates with the Gaussian distributions and the contours using a matplotlib figure.

Results and discussion

The use of similarity maps is demonstrated using ligands of the dopamine D3 receptor. The D3 receptor is one of five subtypes that belong to the G protein-coupled receptor (GPCR) superfamily. D3 receptor ligands contain a positively charged group, usually a protonatable tertiary amine, which forms a structurally and pharmacologically critical salt bridge to the carboxylate of $Asp110^{3.32}$ as found by site-directed mutagenesis [13] and confirmed by the crystal structure [14]. $Asp110^{3.32}$ is highly conserved in all aminergic

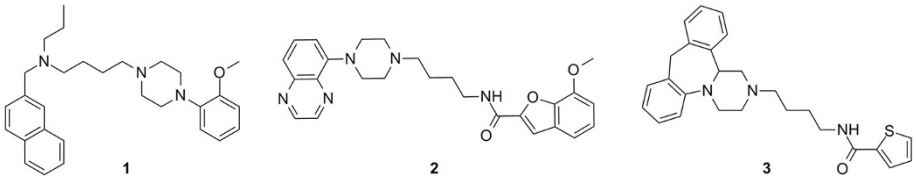

Figure 1 Three dopamine D3 receptor ligands. Reference compound **1** and test molecules **2** and **3**.

receptors. Three active molecules (activity smaller than 10 μM) of the D3 receptor (ChEMBL [15,16] target ID 130) from three different scientific papers [17-19] were extracted from the ChEMBL database (Figure 1). Molecule **1** was selected as reference compound and the other two as test molecules.

Standard fingerprints

The similarity between the reference compound **1** and the test molecules was calculated using four different 2D fingerprints: atom pairs (AP) [20], circular fingerprint [21] with radius 2 as bit vector (Morgan2) and as count vector (CountMorgan2), and feature-based circular fingerprint [21] with radius 2 as bit vector (FeatMorgan2). The fingerprints are described in detail in [22]. Morgan2 is the RDKit implementation of the familiar ECFP4, CountMorgan2 corresponds to ECFC4 and FeatMorgan2 to FCFP4 [23]. The features used by the RDKit for FeatMorgan2 are adapted from [24] and consist of donors, acceptors, aromatic atoms, halogens, basic and acidic atoms. The numerical similarity and maximum differences obtained for the four fingerprints are given in Table 1.

The similarity maps of molecules **2** and **3** using the AP fingerprint are shown in Figure 2. An atom in the AP fingerprint sees all other atoms (if the path is maximum 30 bonds). Atoms with green weights have a majority of paths which are also in the reference compound; deleting them from the fingerprint reduces the similarity to the reference compound. The similarity maps in Figure 2 are consistent with our expectations. For molecule **2**, atoms in the phenyl rings, the piperazine moiety and the alkyl linker were found important for similarity, whereas removing the bits of the nitrogens in the quinoxaline moiety, the oxygen in the benzofuran moiety, or the amide increased the similarity. Also for molecule **3**, atoms in the alkyl linker and partly in the piperazine moiety were found to be most important for similarity.

The similarity maps of the circular fingerprints, Morgan2, CountMorgan2 and FeatMorgan2, are shown in Figure 3. In circular fingerprints, an atom sees only a local environment. Again, the piperazine moiety together with the alkyl linker as well as part of the 7-methoxybenzofuran are highlighted green in molecule **2** for all three variants of the circular fingerprint. Interestingly, the pyrazine part of quinoxaline and the amide appear more pink for CountMorgan2 than for Morgan2. In the first case, one can observe the difference between using a count vector and a bit vector. Using CountMorgan2, the count of the radius-0 bit of the unsubstituted carbons of the pyrazine moiety is 11 for the reference compound and nine for molecule **2**, the count of the radius-1 bit is zero and two. Using Morgan2, the radius-0 bit is set to one in both molecules, whereas the radius-1 bit is zero in the reference compound and one in molecule **2**. Removing the radius-1 bit or decreasing its count will increase the similarity. Removing the radius-0 bit will decrease the similarity, whereas decreasing its count from nine to eight will only have a very small effect on similarity. Thus, the overall "atomic weight" of these carbons is negative (pink) for CountMorgan2, but neutral for Morgan2. The reason for the different appearance of the amide bond, on the other hand, is a hash collision (Figure 4) in the Morgan2 fingerprint: an environment of the amide moiety is hashed to the same bit as a part of the alkyl linker. The

Table 1 Dice similarities and maximum weights

FP	s_{Dice}		Max. Weight		Method	PP		Max. Weight	
	2	**3**	**2**	**3**		**2**	**3**	**2**	**3**
AP	0.604	0.531	0.028	0.033	RF(Morgan2)	0.950	0.600	0.660	0.370
Morgan2	0.561	0.381	0.123	0.110	NB(Morgan2)	1.000	0.999	12.5	20.92
CountMorgan2	0.599	0.529	0.091	0.049					
FeatMorgan2	0.554	0.469	0.176	0.171					

Dice similarity s_{Dice} and maximum weight between reference compound **1** and molecules **2** and **3** using the basic fingerprints (FP) atom pairs (AP), Morgan fingerprint as bit vector with radius 2 (Morgan2), Morgan fingerprint as count vector with radius 2 (CountMorgan2) and Feature Morgan fingerprint with radius 2 (FeatMorgan2). Predicted probability to be active (PP) and maximum weight for random forest (RF) and naïve Bayes (NB) for molecules **2** and **3** with basic fingerprint Morgan2. The bit vectors of the circular fingerprints had the size 1024 bits. The default maximum path length of 30 was used for AP.

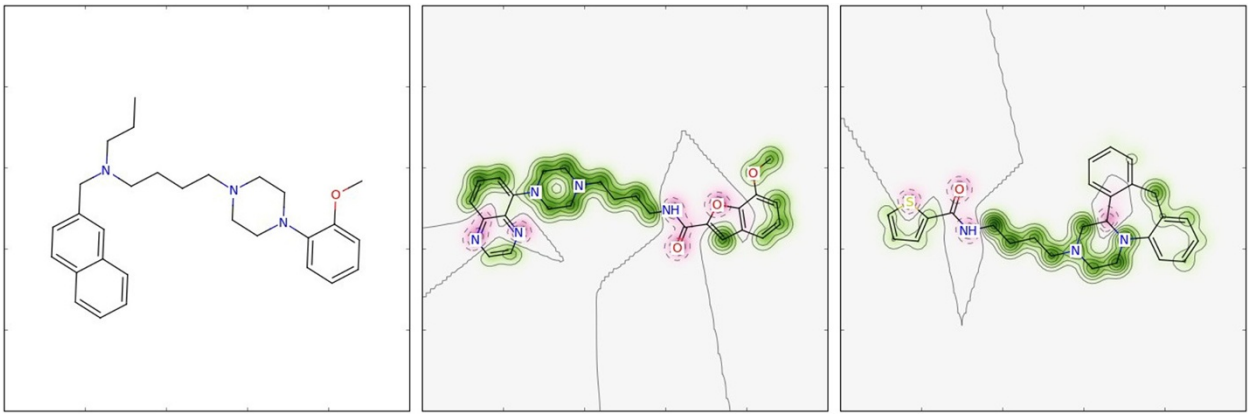

Figure 2 Similarity maps for atom-pairs (AP) fingerprint. Similarity map of molecule **2** (middle) and molecule **3** (right) using AP. The reference compound is molecule **1** (left). Color scheme: removing bits decreases similarity (i.e. positive difference) (green), no change in similarity (gray), removing bits increases similarity (i.e. negative difference) (pink). The default maximum path length of 30 was used for AP.

same effect can be observed for molecule **3**. This collision appears only in Morgan2, which is hashed to a size of 2^{10} bits whereas CountMorgan2 uses 2^{32} bits. It is generally important to use a sufficiently large hash space as collisions can impact the performance of a fingerprint [25]. However, the occurrence of collisions is also dependent on the hashing algorithm used. For Morgan2, increasing the bit-vector size from 2^{10} bits to 2^{14} bits had no influence on the performance [22], and also in the current case doubling the hash space (i.e. 2^{11} bits) did not remove the observed collision (data not shown).

The features in the reference compound are aromatic rings, two acceptors and two basic acceptors. These features are marked green in the right panels in Figure 3 for both molecules. Removing the aromatic acceptors or the donor in the molecules, on the other hand, increased the similarity to the reference compound. Interestingly, one carbon of the piperazine

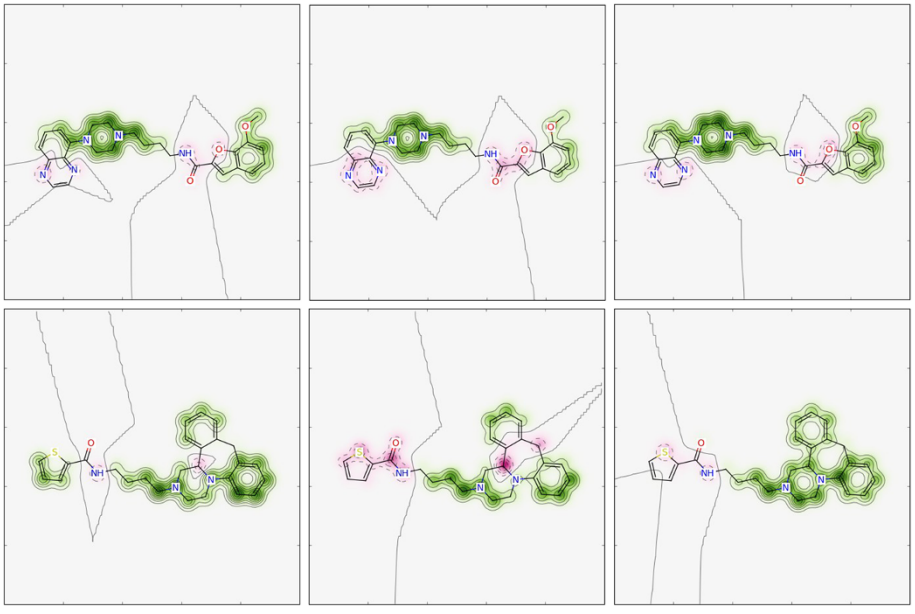

Figure 3 Similarity maps for circular fingerprints. Similarity map of molecule **2** (middle) and molecule **3** (bottom) using Morgan2 (left), CountMorgan2 (middle) and FeatMorgan2 (right). The reference compound is molecule **1** (left panel in Figure 2). Color scheme: removing bits decreases similarity (i.e. positive difference) (green), no change in similarity (gray), removing bits increases similarity (i.e. negative difference) (pink). The bit vectors of the circular fingerprints had the size 1024 bits.

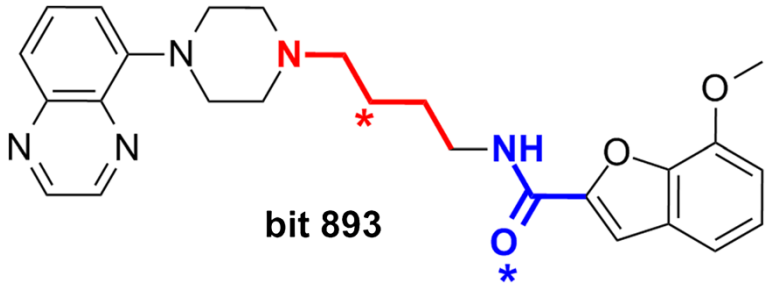

Figure 4 Bit collision in Morgan2/CountMorgan2. Bit Collision in the Morgan2 and CountMorgan2 fingerprint observed for molecule **2** (and analoguously in molecule **3**). The environments highlighted red and blue are hashed to the same bit. The centre atom of an environment is marked with a star.

moiety in molecule **3** is highlighted pink using CountMorgan2 (and to a lesser extent using Morgan2) whereas it is green using FeatMorgan2. For (Count)Morgan2, the atom type of this carbon is different than the atom types of the other carbons as the number of heavy-atom neighbours and the number of hydrogens is different. Using features (donor, acceptor, aromatic, basic, acidic, no-feature), however, the number of neighbours and hydrogens are not considered, thus the feature type (i.e. no-feature) is the same for all carbons in the piperazine.

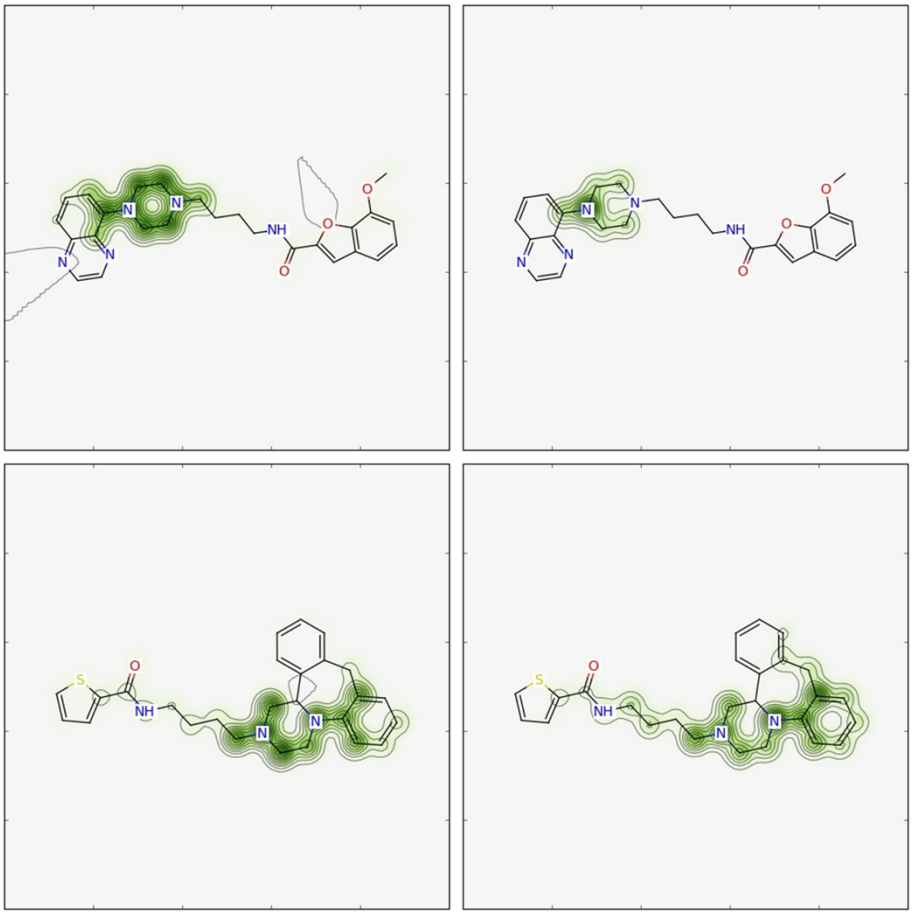

Figure 5 Similarity maps for machine-learning methods. Similarity map of molecule **2** (top) and molecule **3** (bottom) using RF(Morgan2) (left) and NB(Morgan2) (right). Color scheme: removing bits decreases similarity (i.e. positive difference) (green), no change in similarity (gray), removing bits increases similarity (i.e. negative difference) (pink). The bit-vector size of Morgan2 was 1024 bits.

Machine-learning methods

Two kinds of machine-learning (ML) methods, random forest (RF) and naïve Bayes (NB), were trained and used to predict the probability to be active of new molecules. The reference compound and the other active molecules (activity smaller than 10 μM) from Ref. [17] (Figure S1 in Additional file 2) were used together with randomly selected 10% of the 10000 ChEMBL decoys used in a recent benchmarking study [22] to train the ML models. Morgan2 was used as the standard fingerprint. The following optimal parameters of random forests have been determined through a grid search: number of trees (N_T) = 100, maximum depth = 2, minimum samples to split = 2 and minimum samples per leaf = 1. To avoid the problems caused by imbalance in the training set (i.e. many more inactives than actives) for RFs, the balanced random forest algorithm [26] was applied: for each decision tree the majority class is down-sampled to yield an equal number of instances as the minority class. The naïve Bayes classifier was trained using an additive Laplace smoothing parameter of 1.0 and learned class prior probabilities.

The similarity maps (or predicted probability maps, respectively) for the RF model trained with Morgan2 are shown in the left panels of Figure 5. For both molecules, the RF picked up the piperazine moiety with the attached alkyl chain and part of the aromatic fragment. Looking at the active molecules of Ref. [17] (Figure S1 in Additional file 2) confirms that the aromatic ring - piperazine - alkyl chain motif appears in the vast majority of active compounds. Thus, the RF model was able to extract the important structural feature for activity: the nitrogen in the piperazine moiety is protonated at physiological pH and forms the critical salt bridge with Asp110$^{3.32}$ of the receptor [13,14].

Similar findings were obtained for the NB model (right panels in Figure 5). Again, the piperazine moiety was found to be most important.

Conclusions

Similarity maps are an easy and general strategy for the visualization of the atomic origins of fingerprint similarity between molecules. The "atomic weights" are generated by removing the bits belonging to the corresponding atom and comparing the resulting similarity with the similarity of the unmodified fingerprint. Similarity maps can be generated for every fingerprint that allows a backtracking of the bits to a corresponding atom or substructure. The methodology can be extended to machine-learning (ML) models to visualize the atomic contributions to the predicted probability of the ML model. This is especially useful as ML models often appear as black boxes. In future work, we will investigate the application of the visualization strategy to descriptor-based models for physicochemical-property prediction.

Availability and requirements

The source code is provided in Additional file 1. The implementation used the open-source Python toolkits RDKit [10] version 2013.03, scikit-learn [11] version 0.13, and matplotlib [12] version 1.1.0.

Additional files

Additional file 1: Source Code. The file python_scripts.zip contains the source code of the visualization method and the SMILES of the compounds used to generate the figures in the publication.

Additional file 2: Supplementary Figures and Tables. The file supplementary.pdf contains the additional figure mentioned in the text.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SR participated in the conception of the visualization approach, collected the data sets, developed and generated the similarity maps, and drafted the manuscript. GL participated in the conception of the visualization approach and in the discussion of the results, and helped to draft the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

S. R. thanks the Novartis Institutes for BioMedical Research education office for a Presidential Postdoctoral Fellowship. The authors thank Nikolas Fechner for the helpful discussions.

Received: 23 May 2013 Accepted: 23 July 2013
Published: 24 September 2013

References

1. Rogers D, Tanimoto TT: **A computer program for classifying plants.** *Science* 1960, **132:**1115–1118.
2. Dice LR: **Measures of the amount of ecological association between species.** *Ecology* 1945, **26:**297–302.
3. Hansen K, Baehrens D, Schroeter T, Rupp M, Müller KR: **Visual interpretation of kernel-based prediction models.** *Mol Inf* 2011, **30:**817–826.
4. Shemetulskis NE, Weiniger D, Blankey CJ, Yang JJ, Humblet C: **Stigmata: an algorithm to determine structural commonalities in diverse datasets.** *J Chem Inf Comput Sci* 1996, **36:**862–871.
5. Wild DJ, Blankley CJ: **VisualiSAR: a web-based application for clustering, structure browsing, and structure-activity relationship study.** *J Mol Graph Model* 1999, **17:**85–89.
6. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G: **Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors.** *J Med Chem* 2005, **48:**6997–7004.
7. Rosenbaum L, Hinselmann G, Jahn A, Zell A: **Interpreting linear support vector machine models with heat map molecule coloring.** *J Cheminf* 2011, **3:**11–22.
8. Segall M, Champness E, Obrezanova O, Leeding C: **Beyond profiling: using ADMET models to guide decisions.** *Chem Biodivers* 2009, **6:**2144–2151.
9. **Glowing Molecule visualization tool by Optibrium** [http://www.optibrium.com/community/faq/glowing-molecule]
10. **RDKit: Cheminformatics and Machine Learning Software 2013** [http://www.rdkit.org]
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E: **Scikit-learn: Machine Learning in Python.** *J Mach Learn Res* 2011, **12:**2825–2830.

12. Hunter JD: **Matplotlib: a 2D graphics environment.** *Comput Sci Eng* 2007, **9:**90–95.
13. Shi L, Javitch JA: **The binding site of aminergic G protein-coupled receptors: the transmembrane segments and second extracellular loop.** *Annu Rev Pharmacol Toxicol* 2002, **42:**437–467.
14. Chien EY, Liu W, Zhao Q, Katritch V, Han GW, Hanson MA, Shi L, Newman AH, Javitch JA, Cherezov V, Stevens RC: **Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist.** *Science* 2010, **330:**1091–1095.
15. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40:**D1100—D1107.
16. **ChEMBL: European Bioinformatics Institute (EBI), version 14.** Cambridge, UK 2012 [http://www.ebi.ac.uk/chembl/]
17. Banala AK, Levy BA, Khatri SS, Furman CA, Roof RA, Mishra Y, Griffin SA, Sibley DR, Luedtke RR, Newman AH: **N-(3-Fluoro-4-(4-(2-methoxy or 2,3-dichlorophenyl)piperazine-1-yl)butyl)arylcarboxamides as selective dopamine D3 receptor ligands: critical role of the carboxamide linker for D3 receptor selectivity.** *J Med Chem* 2011, **54:**3581–3594.
18. Leopoldo M, Lacivita E, Giorgio PD, Colabufo NA, Niso M, Berardi F, Perrone R: **Design, synthesis, and binding affinities of potential positron emission tomography (PET) ligands for visualization of brain dopamine D_3 receptors.** *J Med Chem* 2006, **49:**358–365.
19. Sasse BC, Mach UR, Leppaenen J, Calmels T, Stark H: **Hybrid approach for the design of highly affine and selective dopamine D_3 receptor ligands using privileged scaffolds of biogenic amine GPCR ligands.** *Bioorg Med Chem* 2007, **15:**7258–7273.
20. Carhart RE, Smith DH, Venkataraghavan R: **Atom pairs as molecular features in structure-activity studies: definition and applications.** *J Chem Inf Comput Sci* 1985, **25:**64–73.
21. Rogers D, Hahn M: **Extended-connectivity fingerprints.** *J Chem Inf Model* 2010, **50:**742–754.
22. Riniker S, Landrum G: **Open source platform to benchmark fingerprints for ligand-based virtual screening.** *J Cheminf* 2013, **5:**26.
23. Landrum G, Lewis R, Palmer A, Stiefl N, Vulpetti A: **Making sure there's a "give" associated with the "take": producing and using open-source software in big pharma.** *J Cheminf* 2011, **3**(Suppl 1):O3.
24. Gobbi A, Poppinger D: **Genetic optimization of combinatorial libraries.** *Biotech Bioeng* 1998, **61:**47–54.
25. Sastry M, Lowrie JF, Dixon SL, Sherman W: **Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments.** *J Chem Inf Model* 2010, **50:** 771–784.
26. Chen C, Liaw A, Breiman L: *Using Random Forest to Learn Imbalanced Data*. Berkeley: University of California; 2004.

doi:10.1186/1758-2946-5-43
Cite this article as: Riniker and Landrum: Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics* 2013 **5**:43.