Identification and analysis of activity cliffs using 3D similarity techniques

Mark D Mackey[*]*, Timothy J Cheeseright, Paolo Tosco.*

Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire SG8 0SS, UK

ABSTRACT

The analysis of activity landscapes and activity cliffs is a widely used method to locate critical regions of SAR. Knowledge of what changes in a series of molecules caused unexpectedly large changes in affinity allows the chemist to focus on the molecular features which are crucial for activity. We examine the usefulness of activity cliff analysis with a metric based on 3D shape and electrostatic similarity, utilizing a ligand-based alignment method. We demonstrate that 3D activity cliff analysis is complementary to the more usual 2D fingerprint-based methods, in that each finds cliffs that the other misses. Moreover, we show that analysis of the activity landscape in the context of a consensus 3D alignment allows the source of the activity cliff to be investigated in terms of the effect that a structural change has on the steric and electrostatic properties of a molecule. The technique is illustrated with two set of compounds with activity against acetylcholinesterase and dipeptidyl peptidase.

INTRODUCTION

The concepts of activity landscapes and activity cliffs[1] have been increasingly used in drug discovery to classify the distribution of biological activities in compound space. Many modelling techniques implicitly assume the similarity hypothesis: for a given similarity metric similar compounds will have similar biological activities, and hence small changes to a molecule should only give small changes in activity. Most of the time this is true, but the cases where the similarity hypothesis breaks down are often the most useful to gain a full understanding of the interactions of a ligand with a target protein.[2,3]

Analysis of the activity landscape is facilitated by examining its slope, i.e. the change in biological activity relative to the amount of structural change. The ratio of these two factors was widely used within Merck in the late 1990s and was termed the "disparity index", but re-emerged later in the literature as the Structure-Activity Landscape Index (SALI).[4] The concept has since been extended in various directions, exploring different descriptors and metrics for similarity,[5] substructure matching and matched molecular pair analysis,[6,7] together with various visual methods to display the activity landscape.[8] A recent review examines the current state of activity cliff analysis and its use in drug discovery.[9]

The most significant variable in analysis of the activity landscape is deciding on its functional form, i.e. what similarity metric is to be employed. It is known that different molecular similarity metrics can have very different neighborhood properties,[1,5] meaning that a pair of compounds forming a significant activity cliff according to one metric may have no significance according to another. The vast majority of work on activity cliff analysis has utilized 2D metrics such as

fingerprints. 2D metrics have significant advantages: they are well-defined, are generally simple and very fast to calculate, and are invariant of 3D shape and conformation. However, ligand binding is an inherently 3D process, so comparison of ligands in 3D should be able to provide information that is not available in 2D. One problem is that 3D similarity is in general not defined on molecules, but on conformations – if a metric is not sensitive to molecular conformation it is not a 3D metric, by definition. For a given pair of molecules, each of which may possess hundreds of energetically-accessible conformations, the concept of 3D similarity becomes highly context-dependent – one conformation of the first molecule may have a matching conformation in the second, but a different conformation may have no such equivalent.

In the context of analyzing an activity landscape, however, there is a solution to this quandary. We are not trying to calculate the similarity of two molecules in a contextual void. Rather, we are interested in their 3D similarity with respect to a biological activity. If we can therefore determine (experimentally or computationally) the bioactive conformation for each molecule, then the 3D similarity metric computed just on those conformations suffices.[10]

Activity cliff analysis has indeed been performed using a 3D similarity metric (a modified shape similarity algorithm) utilizing known bioactive conformations from protein x-ray structures.[10–12] It was found that the analysis using the 3D similarity metric identified substantially different cliffs to an analysis using a 2D fingerprint similarity metric, indicating that significant extra information about the SAR could be obtained from a 3D analysis. Additionally, analysis of identified activity cliffs in the context of the protein active site provides valuable information regarding the source of the observed activity change: does it derive from a change in hydrogen bonding, from lipophilic or aromatic interactions, from changes in bonding to water molecules, from stereochemistry, from steric considerations or from a combination of these effects? [12,13]

A recent study has extended the concept of activity cliff analysis to modelled binding modes using a docking algorithm.[14] It was found that if the bound poses of the active compounds can be reliably determined from an *in silico* protocol, then 3D activity cliff analysis is feasible. Moreover, the structure-based methods can in some cases cope with structural rearrangement of the receptor, which is difficult or impossible with ligand-based methods. This technique greatly extends the utility of 3D activity cliff analysis by removing the requirement for protein crystal structures for every compound in the dataset, which is otherwise extremely limiting. However, it does still require sufficient structural data to perform high-accuracy docking, and this is not always available.

In this paper we present a method of determining 3D activity cliffs in the absence of large amounts of experimental structural information, by utilizing ligand-based alignment techniques. One or more reference molecules are used to provide a conformational context to the data set, and a 3D similarity metric can be applied to the aligned molecules. The use of 3D similarity not only provides a different set of activity cliffs to 2D analysis, but also facilitates the investigation of the causes of any large activity gradients identified in the absence of protein structural information.

METHODS

3D alignment. In order to align a set of active molecules, an initial bioactive conformation of a reference structure needs to be available. This can either be extracted from experimental data or estimated using a pharmacophore generation technique. These involve taking a set of active compounds, computing all possible pairwise alignments over their conformation spaces, and then trying to locate a consensus binding mode which maximizes the similarity of each compound to all of the others. A number of such methods have been described, such as MARS[15], FLAPpharm,[16] and FieldTemplater[17].

Two different alignment methods were utilized in this study to align a set of molecules to a reference bioactive conformation. In the first alignment protocol, conformations for data set compounds other than the reference(s) were generated using the default settings of the XedeX algorithm[18] as implemented in the Forge molecular modelling program.[19] Up to 1000 conformers of each compound were created: the design of the XedeX algorithm ensures that these have a roughly even sampling across the accessible conformation space.

Each compound was then aligned to the reference structure(s) in its data set. The alignments were performed according to a previously described procedure,[20,21] which maximizes the similarity in terms of molecule interaction potentials. In this case the algorithm was modified slightly: the scoring function used was the average of the molecular field similarity[20] and the molecular shape similarity.[22,23] In our experience, using the average of these two similarity metrics consistently gives better alignments than using either alone.

Where a data set possessed multiple reference molecules, the latter were pre-aligned to the correct relative alignment, either by alignment of the α-carbons of the active site where structural data were available, or by using the FieldTemplater algorithm. The alignment of the other compounds to the references was then carried out so as to maximize the sum of the similarity scores to each of the reference molecules simultaneously; i.e., rather than aligning separately to each reference, we align to all of them at once.

The second alignment protocol was designed to reduce alignment noise within closely related series. In this case, a conformation search followed by a free alignment can occasionally lead to misalignment due to undersampling of the conformation space. The solution for structurally related molecules is to recognize that the ideal alignment usually has corresponding atoms placed in close proximity to each other.

First, the maximum common induced subgraph (MCIS) of each molecule with each of the reference molecules was computed.[24] The atoms in the MCIS were directly overlaid onto the corresponding atoms from the reference, provided that chirality and geometry constraints were not violated. The bonds in the MCIS were then frozen, and a constrained conformation search carried out on the remaining bonds only. The resulting conformations were then minimized with the constraints removed, to avoid sterically strained conformers being generated. These conformers were realigned to the reference molecule using a least-squares fit over the MCIS atoms. Finally, the field and shape similarity to the reference molecules was maximized using a rigid-body simplex optimizer. The conformation and orientation with the highest similarity score (averaged over the references) was chosen as the correct alignment for that molecule.

Similarity and Disparity Matrix calculation. Once all molecules in each data set are aligned, a similarity matrix can be computed by taking the aligned conformers in their putative bioactive conformation and computing the combined field and shape similarity score between each pair as detailed above. We originally did so keeping the aligned coordinates fixed; subsequently we found that noise in the data set could be reduced if the conformers in each pair were allowed to relax towards each other slightly. Accordingly, a simplex optimizer was used to maximize the similarity score for each pair by rigid rotation and translation of one conformer onto the other. Only small movements were allowed during this optimization process: on average, molecules move less than 1 Å.

The disparity matrix was computed from the similarity matrix according to the formula

$$D_{ij} = \frac{A_i - A_j}{1 - \min(S_{ij}, 0.95)} \qquad (1)$$

where D_{ij} is the disparity value, A_i is the activity of molecule i (on a log scale), S_{ij} is the similarity value between molecules i and j, and the minimum function is to prevent discontinuities and extreme values at very high similarity.

Activity data usually has some errors associated with it, and it is important to account for this. Two very similar molecules might get a high disparity value from a statistically insignificant activity difference.[25] To avoid this, we specify a minimum activity difference to be treated as meaningful (generally 0.5 log units), and set smaller differences to zero.

Visualisation of the data set. A number of approaches to the visualisation of activity cliff data have been proposed in the literature.[4,26–31] We have found that a combination of local and global views of the disparity matrix data is useful. In addition to viewing the disparity matrix[4] and listing the largest-disparity entries, a local view on one row of the matrix is useful. Given a reference compound (corresponding to one row of the disparity matrix), we display around it the most similar other molecules, along with a graphical representation of the similarity and disparity to the reference. This provides a local view of the activity landscape around one compound; in turn, any of the other molecules can be re-assigned to be the new reference, providing simple navigation across the data set (Figure 1). This method removes the need for arbitrary cutoffs based on activity difference and similarity thresholds to decide whether a molecular pair is an activity cliff or not, and provides a simple visual guide to the local activity landscape.

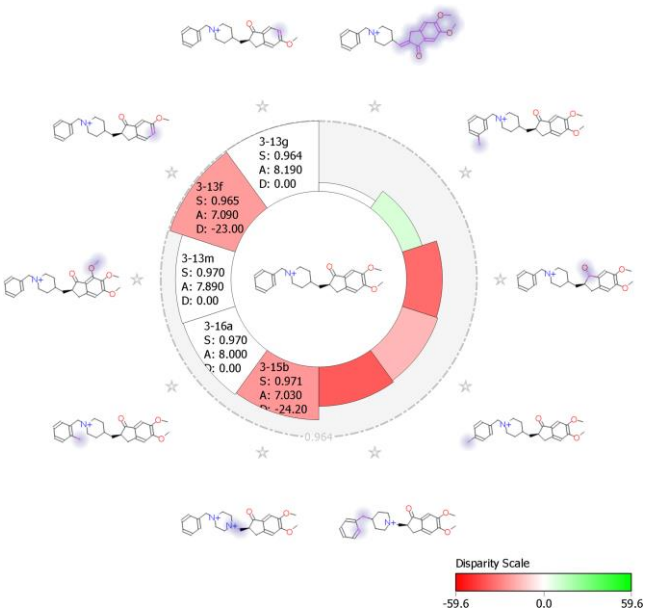

Figure 1. Visualisation of the activity landscape around a molecule. The bar heights represent the distance from the molecule in the center to the one on the edge, while the bars are colored by disparity value. The structural differences between the molecules are highlighted.

Visualisation of activity cliff molecules. A great advantage of 3D activity cliffs is that they are determined from a 3D alignment of two molecules. The fact that an activity cliff exists can thus be augmented by an examination of the differences between the molecules, potentially exploring the reasons for the sudden activity change and increasing understanding of the SAR.

As the molecules in this study are aligned using electrostatic fields and shape, it is instructive to consider the differences between a pair of molecules in terms of shape and electrostatics. While shape is relatively easy to visualize, the change in electrostatics between molecules is more complex. The molecular interaction potentials (MIPs) of the two molecules can be plotted at different contour levels, but for molecules that are highly similar it can be difficult to focus in on the few differences in two complex MIPs. The obvious solution is to contour the difference between the potentials, but a naïve implementation of this leads to plots that are hard to interpret.

The difference map is symmetric, leading to duplication of visual information when the two molecules are displayed side by side.

We circumvent these issues by mostly assigning MIP differences to one molecule only. The algorithm is as follows, for molecules A and B possessing MIPs μ_A and μ_B:

1. Define $\delta = \mu_A - \mu_B$
2. Set δ to zero in regions inside the vdW envelope of either A or B (more specifically, if the vdW contribution to the MIP is positive and larger than the absolute value of the electrostatic contribution)
3. Define $\delta_A = \begin{cases} \delta & if\ sign(\delta\mu_A) > 0 \\ 0 & if\ sign(\delta\mu_A) \leq 0 \end{cases}$
4. Define $\delta_B = \begin{cases} 0 & if\ sign(\delta\mu_B) \geq 0 \\ -\delta & if\ sign(\delta\mu_B) < 0 \end{cases}$

We then plot contours of δ_A and δ_B on A and B respectively. This has the effect that if μ_A and μ_B are both positive, then the contours are only plotted on the one that is more positive, and conversely if both are negative. If they differ in sign, then the relevant contour appears on both. This results in much more intuitive potential difference maps: the contours show which molecule is more positive/negative, not which is less (Figure 2).

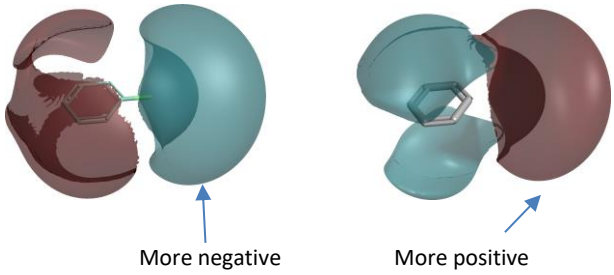

Figure 2. Field difference surfaces for fluorobenzene vs benzene, showing the change in dipole as well as the change in the pi electro density on fluorination

Data Sets. We examine two data sets in this paper, one from the literature and one that we have prepared. The first, a set of 111 compounds active against acetyl cholinesterase (AChE) comes from the well-known Sutherland QSAR data sets.[32] The molecule IDs presented for these molecules are those from the original publication. Molecular field patterns and 3D similarity values were calculated using the pre-existing alignment. 2D similarity values were computed using ECFP4 circular fingerprints and a Tanimoto metric.

The second data set we examined was a set of molecules active against DPP IV. The initial DPP IV data set was obtained via searching BindingDB[33] for compounds with measured activity against DPP IV. The bindingDB ontology has multiple entries for DPP4, so multiple searches were performed and aggregated. The results were filtered to remove activity measurements against non-human species, and processed to remove salts and aggregates multiple activity values against the same structures. The KNIME[34] protocol to perform the filtering is available in the Supporting Information.

The final data set was then manually filtered and checked against the original literature.[35–38] In a number of cases the chirality of the compound was not clear (either because it was synthesized and tested as a racemate or because chirality assignment was not fully performed). The majority of these involved compounds from the triazolopiperazine series from Kim et al.[35] where the chirality of the 8 substituent was not determined. Following the suggestion from that paper we assigned the more active diastereomer as having the 8-substituent congruent to the (R) configuration of compound **46b** in that paper.

Compound 70 in the data set in was chosen as a reference and was aligned to the ligand from PDB entry 1x70. The position of the fluorophenyl substituent was chosen by reference to the crystal structure shown in Kim et al.[35] This reference proved to perform poorly in aligning the

compounds in the data set with an opposite configuration at the 8-position, such as compound 71. Accordingly, the piperazine ring conformation and the orientation of the fluorophenyl substituent in compound 71 were manually adjusted into a reasonable conformation. Compound 71 was then added as an additional reference.

Compounds with no substituent at the 8 position, or whose configuration at that position matched compound 70, were aligned to compound 70 using the substructure alignment algorithm in Forge[19] with "Normal" settings. The remaining compounds were aligned to both compounds 70 and 71. The aligned project is attached in the Supporting Information.

RESULTS AND DISCUSSION

AChE data set. The AChE data set comprises 111 molecules, with a pK_i range from 4.27 to 9.52. Figure 3 shows a comparison between the 2D and 3D similarity metrics on this data set. As can be seen, the scale of the two similarity metrics is rather different: only a small proportion of molecule pairs in this data set have an ECFP4 Tanimoto score of above 0.7, while the 3D similarity values are generally higher. The scatter plot shows that (as expected) there is some correlation between 2D and 3D similarity, but it is fairly low (r^2=0.51). In particular, there are numerous points where the 3D similarity is very high while the 2D similarity is low. We would thus expect to see significant differences in which pairs of compounds are found to have high disparity values between the two methods: the question is whether one provides more useful information than the other.

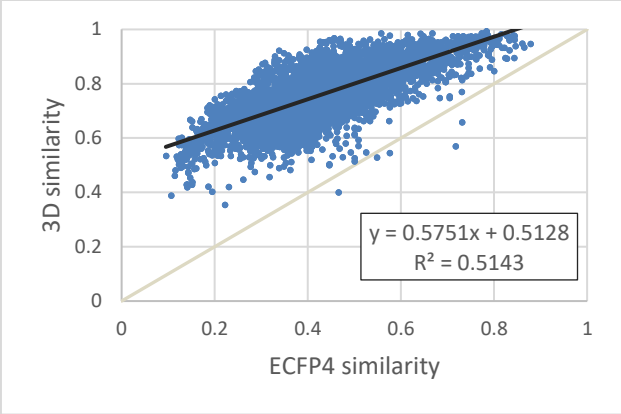

Figure 3. Comparison of 2D and 3D similarity across the AChE data set

The 2D and 3D disparity matrices were inspected to find molecule pairs that differ greatly in 2D and 3D disparity. Comparison of these showed some significant differences between the types of activity cliff that are able to be identified by the two methods. Molecules 2-27, 2-34 and 2-35 provide a clear example (Figure 4). Molecules 2-34 and 2-35 form an activity cliff in both 2D and 3D – the difference from cyclohexyl to cyclopropyl is relatively minor in both similarity metrics, while the cyclohexyl compound is significantly more active. However, comparing a cyclohexyl substituent to the *p*-toluyl substituent in 2-27, the 2D similarity is relatively low (0.654), while the 3D shape and electrostatic similarity is still quite high (0.944). Although the toluyl group has more electrostatic character than the cyclohexyl, this is relatively weak, and the shape and hydrophobicity match between the two is very high. The 2D similarity metric treats aromatic carbons as being totally dissimilar to aliphatic ones, so misses this activity cliff. In this case the 3D analysis finds important SAR that is missed by the fingerprints.

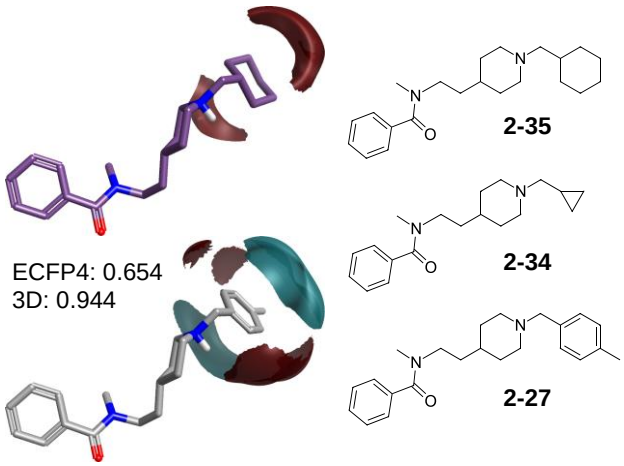

Figure 4. Changing aliphatic rings to aromatic is a larger change in 2D than in 3D

Molecules 1-11 and 1-13 provide a counter example, where activity cliffs are seen in 2D but not in 3D. These molecule differ only in a chain extension (Figure 5). Their 2D similarity is relatively high (0.72), and they form an activity cliff according to the definition in Hu and Bajorath.[10] However, as is apparent from Figure 6, the two molecules have a poor 3D alignment and hence a low 3D similarity. The 3D similarity function is particularly sensitive to the chain length change as is has a large effect on the conformation of the molecule.

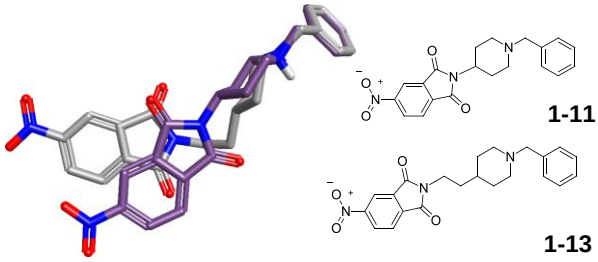

Figure 5. Chain extensions form an activity cliff in 2D, but not in 3D

As well as chain insertions/deletions, this effect will also show up for enantiomeric pairs (where the 2D similarity is 1 by most metrics) and for changes where the substitution greatly affects the accessible conformation space of the molecule, such as *ortho* substitutions on aromatic systems.

In these cases the large conformational difference between the molecule leads to a low 3D similarity value and hence a low disparity.

One effect that was noted in the fingerprint analysis of the AChE compounds was that the effect of a change on the similarity score can depend dramatically on where the change occurs. Altering the center of a molecule causes a much larger similarity drop than altering the edge, at least with the circular fingerprints. An example is provided by the reduction of the carbonyl group from molecule 3-13e to molecule 3-17 and concomitant drop in activity of 1.7 log units. Because this relatively minor structural change occurs near the center of the molecule and involves a hybridization change, the ECFP4 similarity is only 0.593, and no activity cliff is detected. However, the –OH group and the carbonyl oxygen are both electronegative and both have H-bond acceptor functionality. The difference in the MIPs of the two molecules is relatively modest, leading to a 3D activity cliff being detected (Figure 6).

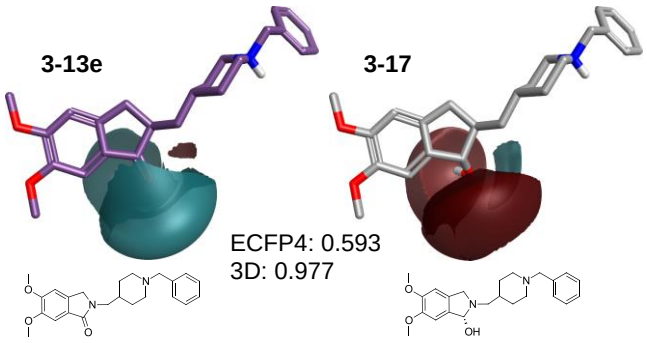

Figure 6. A hybridization change in the center of a molecule reduces 2D similarity so that an activity cliff is not seen in 2D

Viewing the local activity landscape around a particular active compound can provide additional information that is not available just from consideration of the activity cliffs. For example, Figure 1 shows the local activity landscape around compound 3-13e. Four activity cliffs (disparity>20)

are immediately apparent (the dark red wedges in Figure 2). One is reduction of the amide as shown in Figure 6 and discussed earlier. Two cliffs involve reversal of the central piperidine and conversion of the piperidine to a piperazine respectively. The fact that both of these changes cause a large activity drop indicates that the introduction of a basic nitrogen adjacent to the cyclic ketone is as much a cause of a drop in activity as the removal of the basic nitrogen at the other end of the ring.

The fourth activity cliff identified is removal of the 5-methoxy group. The advantage of being able to compare all close neighbors, not just the cliffs, is evident here as it can be seen that although removal of the 5-methoxy group causes a large drop in activity, the 4-methoxy group can be removed at no penalty. Examination of the other neighbors of 3-13e shows that methylation of the benzyl group is allowed or favorable at the 2- and 3-positions but disfavored at the 4-position, hinting at a steric constraint.

The activity view as shown in Figure 1 thus combines very well with the activity cliff analyses. After identification of a cliff, the local activity landscape around the cliff compounds can be compared to identify what other small changes have been made to either compound. The effect of these changes on activity provides extra information that assists in the interpretation of the source of the cliff.

DPP IV data set. The DPP IV data set consists of 91 molecules with a pK_i range from 5.7 to 9.7. Analysis of these compounds using 3D shape/field similarity and 2D fingerprint similarity shows distinct differences in the activity cliffs detected. The most striking of these involves the diastereomeric pairs within the data set. Inversion at the 8-position of the triazolopiperazine results in a large loss of activity. Since 2D fingerprint metrics are insensitive to chirality, these diastereomeric pairs have a similarity of 1, leading to large 2D activity cliffs being detected. For

example, compounds 70 and 71 in the data set are diastereomers, with activities 9.74 and 6.94 respectively (Figure 7). The diastereomers have quite different modelled bound conformations so their 3D similarity is low (0.79). As a result this pair does not form a significant activity cliff in 3D.

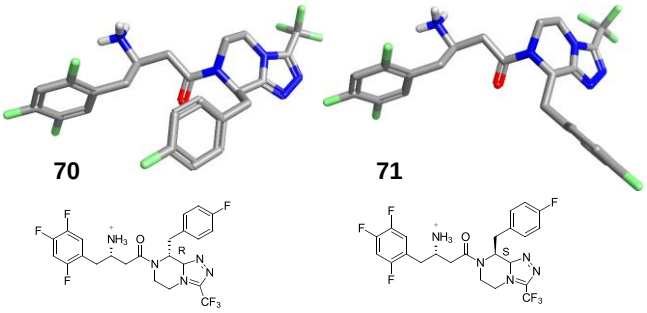

Figure 7. Diastereomeric pairs form a cliff in 2D but not in 3D

These diastereomeric differences dominate the 2D analysis of this data set: the largest cliffs are all between diastereomeric pairs, and even if these are removed the next set of large cliffs are all between molecules with different stereochemistry at the 8 position. A useful analysis using 2D fingerprints was only possible by filtering out all of the compounds having one configuration (the less-active *S* diastereomer in Figure 8) and examining the remainder. One of the largest non-diastereomeric 2D cliffs in the data set was between molecules 71 and 80, with the replacement of a 4-fluorobenzyl by a 3,5-di-trifluoromethylbenzyl group, with an ECFP4 similarity of 0.81. When compared in terms of 3D shape and electrostatics, the large difference in shape between the two substituted phenyls means that this pair does not have an exceptional disparity value (Figure 8).

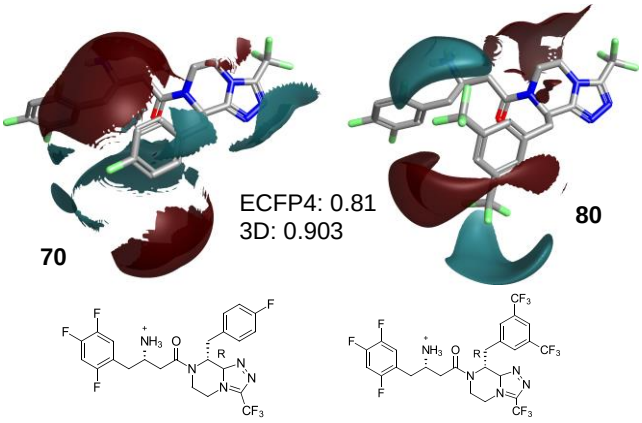

Figure 8. Large electronegativity changes cause a 2D cliff to not appear as significant in 3D

In contrast, molecules 28 and 82 have two minor differences: the removal of a methyl group from the triazolopiperazine ring and the moving of a fluorine on the benzyl group from the 4- to the 3-position, which together lead to a change in activity of more than two log units. Using the ECFP4 metric, these two changes are sufficient to render these molecules quite different (similarity 0.493), while in 3D the molecules are still seen as very similar (0.962) and hence a large disparity value is computed (Figure 9).

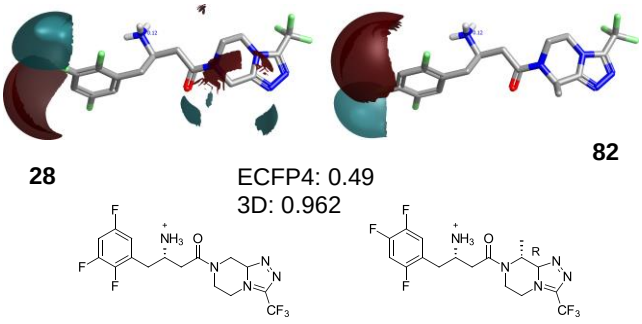

Figure 9. Multiple changes prevent activity cliff detection in 2D

This shows a general feature of many 2D similarity metrics, and the circular fingerprint ones in particular. Applying more than one small change to a molecule usually reduces the similarity by more than applying one larger change: adding an ethyl group to a molecule is a much "smaller" change than adding two separate methyl groups in different locations. As a result, 2D activity cliff

identification is generally restricted to single point changes. In contrast, the 3D shape and electrostatic similarity metrics are capable of locating interesting pairs of molecules that differ by two or three small changes, which can help locate cooperative or nonlinear effects of different substitutions. In many cases not all combinations of substituents are synthesized, and so effectively restricting the analysis to single-point changes (via use of a particular similarity metric or by performing the analysis only on matched molecular pairs) important information may be lost.

A large number of the high-disparity pairs in the DPP IV data set involve changes to the halogen substitution pattern. Changing the number, position and element of the halogens can cause more than two log units of activity change. Examination of the MIP difference maps across multiple such pairs shows that the effects of halogen substitution are nonlocal and often subtle. For example, adding a fluorine not only increases the negative electrostatic potential near that fluorine: it also changes the dipole moment across the ring, as well as withdrawing electron density from it making the ring a stronger π acceptor (Figure 10). From the crystal structure, this ring stacks against the face of Tyr662, so it is likely that some of the increased affinity of **59** vs **15** can be attributed to the increased strength of this stacking due to electron withdrawal by the extra fluorine.

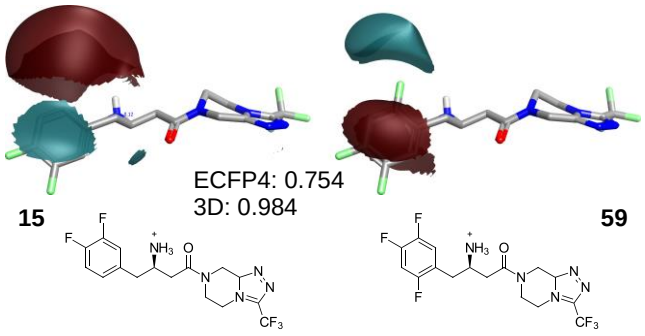

Figure 10. Halogen substitution has nonlocal effects

Figure 11 shows the activity landscape around molecule 59. Several features immediately become apparent from this view. The first is that methyl substitution at the 8-position of the

imidazopyridazine is beneficial for activity, but only if the stereochemistry is correct. The second is that the halogen substitution pattern on the phenyl group is optimal: any changes seem to decrease activity. In particular, increasing the size of the 4-substituent is penalized, with a strong activity gradient F > Cl >> SMe. The third is that replacing the CF_3 group with other small groups has only a small effect on affinity. All of these immediate conclusions are strongly in agreement with the conclusions reached in the original literature on this series.[35–38]

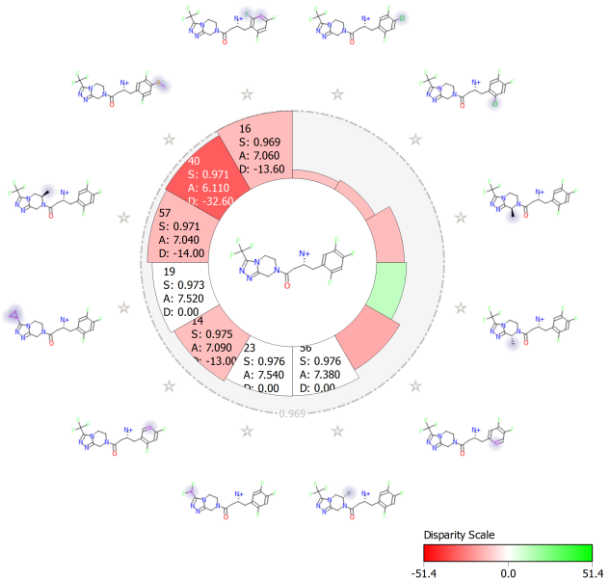

Figure 11. Activity landscape around molecule 59.

CONCLUSIONS

We have demonstrated that disparity analysis in 3D using modelled alignments of ligands is a useful addition to the computational chemist's arsenal of SAR analysis techniques. Care is required in generating the alignments: as in 3D-QSAR techniques, all of the signal is contained in the alignment process and so careful attention to this is required. However, activity cliff analysis is

generally performed within a series, and modern alignment techniques are sufficiently accurate to provide a robust signal-to-noise ratio under such conditions.

Analysis of disparity values in 3D identifies activity cliffs that are not found in a 2D analysis, and hence adds additional information. Since the converse is also true, especially when focus on stereochemistry is desirable, the combination of the two techniques appears to be ideal to detect activity cliffs of different types. One attractive feature of the 3D similarity method used here is that it appears to cope well with molecules that differ through small changes at multiple positions around the molecule. Many 2D similarity techniques, and in particular the circular fingerprints, give very low similarity values one molecules differ in more than one place and hence miss potentially-useful information in the common case where not all combinations of substituents have been synthesized.

Once a 3D activity cliff is identified, examination of the MIP differences between the relevant molecules can provide guidance as to the possible causes of the activity change. In some cases this is obvious, but in others the MIP deltas show not only the direct effects of the structural change but also the indirect effects such as changes in the molecular dipole and the π cloud density on aromatic rings. This information, in turn, allows the modeler to suggest changes to the structure to further improve the activity. The combination of activity cliff analysis and MIP difference visualisation is particularly powerful in that the former flags up the most interesting or important changes in a series, while the latter then assists in explaining the potential sources of the activity change in an easily-understood manner.

All algorithms and visualisation techniques presented herein have been implemented in Forge's Activity Miner module.

ASSOCIATED CONTENT

Supporting Information. The AChE data set as a Forge project file; a KNIME[34] protocol used to process the original DPP IV data set from BindingDB[33]; the final aligned DPP IV data set in SDF format and as a Forge project file.

AUTHOR INFORMATION

Corresponding Author

*Email: mark@cresset-group.com

Author Contributions

MM, TC and PT developed the methods. TC processed the DPP IV data set. MM wrote the manuscript. All authors have given approval to the final version of the manuscript.

Notes

The authors declare a financial interest in that they are employees of and shareholders in Cresset.

ACKNOWLEDGMENTS

Assistance in collating and aligning the DPP IV data set from Giovanna Tedesco is gratefully acknowledged by the authors. We would also like to thank Nigel Palmer for assistance in developing the visualisation algorithms, and Martin Slater, Susana Tomásio and Rob Scoffin for invaluable feedback.

ABBREVIATIONS

DPP IV, dipeptidyl peptidase IV; MCIS, maximum common induced subgraph; SAR, structure-activity relationship; MIP, molecular interaction potential; AChE, acetylcholine esterase

REFERENCES

(1) Maggiora, G. M. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535–1535.
(2) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. *J. Med. Chem.* **2014**, *57* (8), 3186–3204.
(3) Kubinyi, H. *Perspect. Drug Discov. Des.* **1998**, *9-11* (0), 225–252.
(4) Guha, R.; Van Drie, J. H. *J. Chem. Inf. Model.* **2008**, *48* (3), 646–658.
(5) Dimova, D.; Stumpfe, D.; Bajorath, J. *J. Chem. Inf. Model.* **2013**, *53* (9), 2275–2281.
(6) de la Vega de León, A.; Bajorath, J. *J. Chem. Inf. Model.* **2014**, *54* (10), 2654–2663.
(7) Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. *J. Chem. Inf. Model.* **2013**, *53* (7), 1576–1588.
(8) Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. *J. Med. Chem.* **2012**, *55* (11), 5546–5553.
(9) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M.; Borges, F. *Drug Discov. Today* **2014**.
(10) Hu, Y.; Bajorath, J. *J Med Chem* **2012**, *52* (3), 670–677.
(11) Furtmann, N.; Hu, Y.; Bajorath, J. *J. Med. Chem.* **2014**, 140731160933003.
(12) Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. *J. Chem. Inf. Model.* **2012**, *52* (6), 1490–1498.
(13) Seebeck, B.; Wagener, M.; Rarey, M. *ChemMedChem* **2011**, *6* (9), 1630–1639.
(14) Husby, J.; Bottegoni, G.; Kufareva, I.; Abagyan, R.; Cavalli, A. *J. Chem. Inf. Model.* **2015**.
(15) Klabunde, T.; Giegerich, C.; Evers, A. *J. Chem. Inf. Model.* **2012**, *52* (8), 2022–2030.
(16) Cross, S.; Baroni, M.; Goracci, L.; Cruciani, G. *J. Chem. Inf. Model.* **2012**, *52* (10), 2587–2598.
(17) FieldTemplater http://www.cresset-group.com/products/forge/fieldtemplater/ (accessed May 19, 2015).
(18) XedTools http://www.cresset-group.com/products/xedtools/ (accessed Oct 23, 2014).
(19) Forge: Powerful computational tool to understand SAR & design http://www.cresset-group.com/products/forge/ (accessed Oct 23, 2014).
(20) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. *J. Chem. Inf. Model.* **2006**, *46* (2), 665–676.
(21) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. *J. Chem. Inf. Model.* **2008**, *48* (11), 2108–2117.
(22) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99* (11), 3503–3510.
(23) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.
(24) Hariharan, R.; Janakiraman, A.; Nilakantan, R.; Singh, B.; Varghese, S.; Landrum, G.; Schuffenhauer, A. *J. Chem. Inf. Model.* **2011**, *51* (4), 788–806.
(25) Medina-Franco, J. L. *Chem. Biol. Drug Des.* **2013**, *81* (5), 553–556.
(26) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51* (2), 258–266.
(27) Wawer, M.; Bajorath, J. *J. Med. Chem.* **2011**, *54* (8), 2944–2951.
(28) Kayastha, S.; Dimova, D.; Iyer, P.; Vogt, M.; Bajorath, J. *J. Chem. Inf. Model.* **2014**, *54* (2), 442–450.
(29) Iyer, P.; Stumpfe, D.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51* (6), 1281–1286.
(30) Medina-Franco, J. L. *J. Chem. Inf. Model.* **2012**, *52* (10), 2485–2493.
(31) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. *J. Med. Chem.* **2009**, *52* (10), 3212–3224.
(32) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1906–1915.

(33) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.* **2007**, *35* (Database), D198–D201.
(34) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.
(35) Kim, D.; Kowalchick, J. E.; Brockunier, L. L.; Parmee, E. R.; Eiermann, G. J.; Fisher, M. H.; He, H.; Leiting, B.; Lyons, K.; Scapin, G.; Patel, S. B.; Petrov, A.; Pryor, K. D.; Roy, R. S.; Wu, J. K.; Zhang, X.; Wyvratt, M. J.; Zhang, B. B.; Zhu, L.; Thornberry, N. A.; Weber, A. E. *J. Med. Chem.* **2008**, *51* (3), 589–602.
(36) Kim, D.; Wang, L.; Beconi, M.; Eiermann, G. J.; Fisher, M. H.; He, H.; Hickey, G. J.; Kowalchick, J. E.; Leiting, B.; Lyons, K.; others. *J. Med. Chem.* **2005**, *48* (1), 141–151.
(37) Kim, D.; Kowalchick, J. E.; Edmondson, S. D.; Mastracchio, A.; Xu, J.; Eiermann, G. J.; Leiting, B.; Wu, J. K.; Pryor, K. D.; Patel, R. A.; He, H.; Lyons, K. A.; Thornberry, N. A.; Weber, A. E. *Bioorg. Med. Chem. Lett.* **2007**, *17* (12), 3373–3377.
(38) Kowalchick, J. E.; Leiting, B.; Pryor, K. D.; Marsilio, F.; Wu, J. K.; He, H.; Lyons, K. A.; Eiermann, G. J.; Petrov, A.; Scapin, G.; Patel, R. A.; Thornberry, N. A.; Weber, A. E.; Kim, D. *Bioorg. Med. Chem. Lett.* **2007**, *17* (21), 5934–5939.
(39) Czechtizky, W.; Dedio, J.; Desai, B.; Dixon, K.; Farrant, E.; Feng, Q.; Morgan, T.; Parry, D. M.; Ramjee, M. K.; Selway, C. N.; Schmidt, T.; Tarver, G. J.; Wright, A. G. *ACS Med. Chem. Lett.* **2013**, *4* (8), 768–772.

For Table of Contents Use Only

Identification and analysis of activity cliffs using 3D similarity techniques

Mark D Mackey*, Timothy J Cheeseright, Paolo Tosco.