Practical applications of Matched Series Analysis: SAR transfer, binding mode suggestion, and data point validation

Peter Hunt[*1], Matthew Segall[1], Noel O' Boyle[2], Roger Sayle[2],*

1. Optibrium Ltd, 7221 Cambridge Research Park, Beach Drive, Cambridge, CB25 9TL, UK. +44-(0)1223-815900. peter@optibrium.com, matt@optibrium.com

2. NextMove Software Ltd, Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge, CB4 0EY, UK. noel@nextmovesoftware.com, roger@nextmovesoftware.com

Abstract

Background

The assumption in scaffold-hopping is that changing the scaffold does not change the binding mode and the same structure-activity relationships (SAR) are seen for substituents decorating each scaffold.

Results/Methodology

We present the use of Matched Series Analysis, an extension of Matched Molecular Pair Analysis, to automate the analysis of a project's data and detect the presence or absence of comparable SAR between chemical series.

Conclusions

The presence of SAR transfer can confirm the perceived binding mode overlay of different chemotypes or suggest new arrangements between scaffolds that may have gone unnoticed. The absence of series correlation can highlight the presence of inconsistent data points where assay values should be reconfirmed, or provide challenge to any project dogma.

Executive Summary

- Matched Series Analysis extends the concept of Matched Molecular Pair Analysis to consider trends in activities for series of compounds that are identical except for small changes at a single point of substitution.
- Comparing matched series in a project data set with those in a large database of previously observed series, improves the ability to predict new substitutions that are likely to improve target activity.
- Strong correlations between matched series with different scaffolds within a project data set can increase confidence in the hypothesis that they bind in similar orientations.
- A poor correlation between matched series with different scaffolds within a project data set may indicate an alternative binding mode and challenge project assumptions.
- Poor correlations between matched series within a project data set may also indicate outlier data points that should be considered carefully and may be valuable to reconfirm experimentally.

Introduction

Matched Molecular Pairs Analysis (MMPA) [1] is the examination of pairs of molecules that are identical except for replacement of a single, small contiguous fragment (either a substituent or a scaffold) at the same position in each molecule. Matched Series Analysis (MSA) is the extension of MMPA to consider a series of variations at a particular site on a molecule. The variations at this site are ordered with respect to the desired endpoint (usually activity) and then this order is compared to other examples of that series of variations found in the same or other reference datasets, across different scaffolds. MSA was introduced in 2011 by Wawer and Bajorath [2] and the concept extended by O'Boyle *et al.* [3] with a statistical approach to predict the R-groups that would be most likely to improve the activity of a molecule given a previously observed order for the activity of other derivatives.

MSA can be performed via two methods [3] referred to herein as 'SAR transfer' and 'Matsy' respectively. A major purpose of these methods is to find related matched series in large databases in order to identify new opportunities for compound synthesis. The Matsy approach seeks to find many examples of a particular series to build confidence that the suggestions for new compounds (which extend the series) are likely to improve activity over those already investigated. To find sufficient numbers of examples, the matched series tend to be short, typically comprising 3-6 compounds, in length. This contrasts with the SAR transfer method which seeks to find matched series of compounds where the activity orders are highly correlated and thus provide confidence in the new compound suggestions. To ensure that these correlations are meaningful, the matched series considered are much longer (usually >6 compounds) and, as a consequence, there will be fewer examples of these long series in the databases.

The previous uses of the MSA methodology have primarily been to derive information from large external databases, containing up to millions of compounds, to find statistically significant trends. In this paper, we will present examples of the application of the SAR transfer methodology for more qualitative analysis of smaller project data sets (containing hundreds to low thousands of compounds) in a variety of ways. While the correlations found in smaller data sets may not be statistically significant, they can indicate interesting trends or outliers that are worthy of further investigation to confirm or refute hypotheses.

MSA has been shown [4] to be able to identify 'holes' in the SAR where the crossover of SAR is possible and a chemist can transfer the learnings from one series onto another, thus ensuring that opportunities for improved compounds are not missed in the wealth of data accumulated on a project.

The ability to examine these SAR crossovers may, however, also provide evidence of cases where the assumed correspondence of functionality or sidechains has broken down. These examples challenge any dogma that may have arisen during the development of a series, which might be hard to counter with individual matched pair results. The SAR transfer method can also provide evidence for novel hypotheses concerning the overlay of different chemotypes by demonstrating consistent SAR in particular substituents. Finally, we show how the methodology can be used to highlight inconsistent and potentially incorrect assay values that have the potential to misdirect a project if taken out of the context of a series.

Methods

The implementation of the matched series algorithm employed herein is described in detail in O'Boyle *et al.* [3]. Briefly, all matched series in these data were calculated using the method of Hussain and Rea [5]. The fragmentation scheme used involved a single cut at each acyclic single bond in turn if either end of the bond was involved in a ring or if the bond was between a non-sp2-hybridized carbon atom and a non-carbon atom. Scaffolds were required to have 5 or more heavy atoms, while R groups were required to have 12 or fewer heavy atoms (other implementations, with small variations, are available and the reader

is directed to these references for further information on those [6,7]). The matched series analyses and resulting visualisations, described in this paper, were performed with the StarDrop software [8].

A query matched series is identified from the input data and ordered by the observed activity values (top row of Figure 1), correlated series of the same derivatives from other chemotypes is found within the input data and aligned to the query sequence (rows 2 & 3 in Figure 1); the correlation of the matched series activities with the query series is calculated using a Spearman's rank correlation coefficient. If this coefficient is greater than a user-defined cut-off, the matched series is examined to determine if there is a further member of the series with higher activity than all the preceding members. If this is so, this derivative is suggested as a new derivative to be made in the query series (for example the iodine derivative as the left-hand column in Figure 1). As can be seen by the blank cells in rows 2 & 3 of Figure 1, the correlated series does not have to be complete, but simply longer than a user-defined minimum and whilst the target column in Figure 1 is useful for large database searches, the usage described herein is searching a smaller project data set, only coming from a single target and so is not reported.

This SAR transfer method was used to analyse two data sets; a data set of compounds with inhibitory activities measured against the peroxisome proliferator-activated receptor gamma (PPARγ) target and a data set of compounds with inhibitory activities measured against Human cyclin dependent kinase 2 (CDK2). Both sets contain public domain data collected from the ChEMBL database [9] version 20 and are provided in the Supporting Information. The PPARγ set comprises 666 compounds with IC_{50} data in μM, which was converted into pIC_{50} values (the negative log of the IC_{50} in Molar concentration), whilst the CDK2 set has 4536 compounds with IC_{50} values in nM before conversion to pIC_{50}. These data sets contain a range of chemotypes, as illustrated by the structural clustering for the PPARγ set shown in Figure 2.

The PPARγ data set was analysed looking for matched series with a length of 3 or more, in which the PPARγ activity increased with a minimum correlation of 0.2 between each series. This resulted in the generation of a set of 318 suggestions derived from 118 matched series. These parameters were chosen because the set is relatively small and would be unlikely to contain very long matched series. Furthermore, the analysis is directed to identify not only the correlated, successful, SAR transfer but also the unsuccessful, uncorrelated SAR [10]. The CDK2 set gave 2761 suggestions from 618 matched series, based upon MSA with a minimum series length of 3 and a minimum correlation of 0.9. The increase in the correlation parameter was used to focus only on the consistent SAR between series that would be expected if the binding poses of different chemotypes were the same.

Scaffold	Target	Correlation	I	O	Cl	F	F F F	N	O S O		H	O O
				7.721	7.678	7.444	7.18	6.959	5.842	5.833	5.699	5
	Platelet-derived g...	0.8333	6.54	5.27	5.96	5.17		6.07	5.03	5	4.95	4.52
	Aldo-keto-reduct...	0.7381	7.17	6.46	6.96	6.3	7.17		5.26	6.44	5.21	4.52

Figure 1. Three matched series of derivatives in tabular form with a query series from the input data set shown in the top row and two matching series shown in rows 2 and 3. The query series is ordered from right to left in increasing activity and the corresponding activities for the matched series derivatives are shown in the appropriate columns. The colours of the cells are determined by the activity values in each row (red = lowest row value, yellow = highest row value) to visually emphasise the correlations between the query and matched series, quantified by a Spearman's rank correlation coefficient.

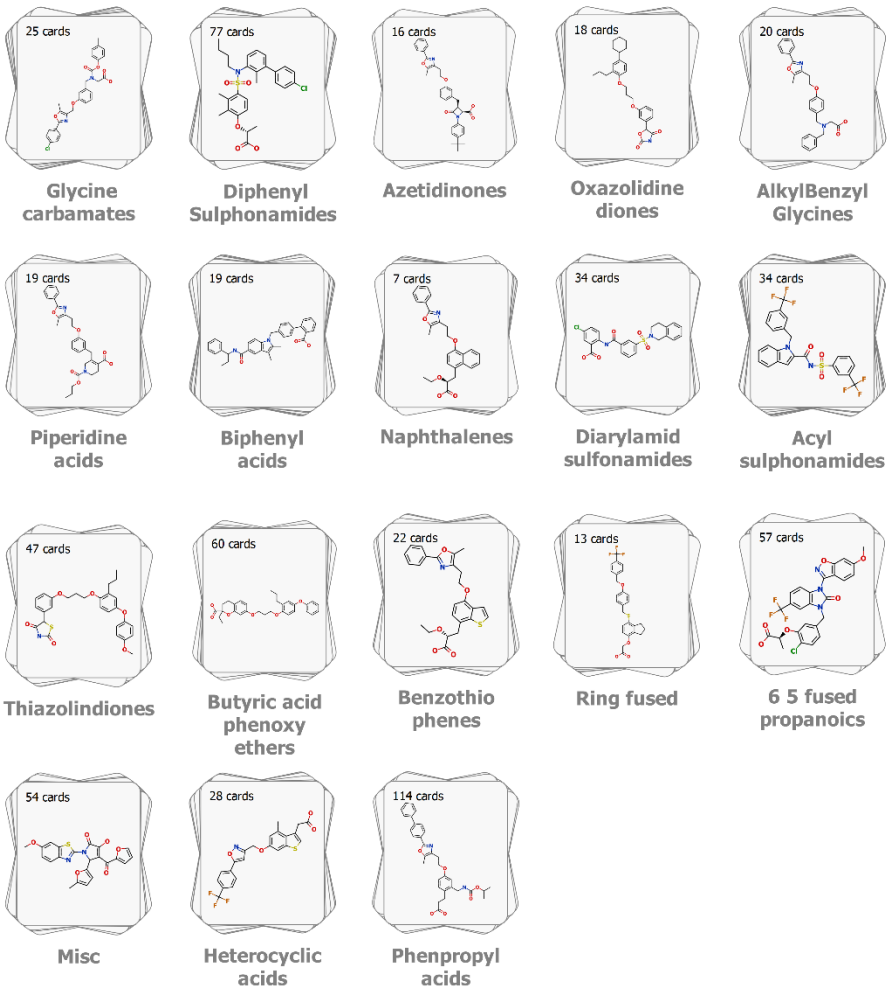


Figure 2. Stacks of cards representing the different subsets of compounds within the PPARγ data set, produced by manual substructure searching and grouping into stacks. On each stack the structure of a representative compound from within the stack and the total number of compounds in the stack are shown. The text label is a general description of the compounds within of the stack.

The output of each analysis is a dataset of new compound suggestions that are predicted to be more active than the most active in the current matched series and the compounds from the input dataset which make up that series. In this case, the use of a short series length makes SAR transfer seem a little like Matsy, and indeed, if the minimum number of observed series for the Matsy methodology was reduced to 2 then some of the suggested molecules would be derived from both methods. The benefit of the SAR transfer methodology is that the correlation of one series with another does not have to be perfect before a comparison or suggestion is proposed. Hence only the SAR transfer methodology is performed in these cases. The interpretation of the resulting output will be covered in the next section.

Results and Discussion

1) New compound suggestions (missed opportunities & hole filling)

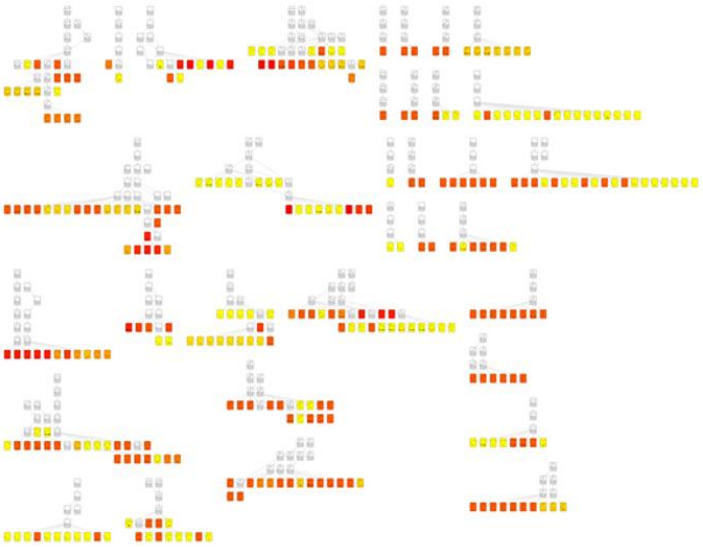
An overview of the MSA output from the analysis of the PPARγ data set is shown in Figure 3(a), where each compound in the output data set is represented by a card [11]. Links between cards indicate the matched series in the data set that, in turn, lead to the novel suggested derivatives, resulting in a network of cards that illustrate the relationships in the data set. The novel compounds are coloured by the different matched series that inspired them and the white cards are the compounds (from the original data set) that make up those matched series. The most striking feature of these results is that, even in this representation of a small project data set, there are plenty of opportunities to explore new compounds that are likely to have improved activity against PPARγ, based upon the SAR already gathered by the project.

Another feature of the output, illustrated in Figure 3(b), is that the correlation in SAR is not constant for any particular network and possible reasons for this will be discussed later. One can examine each network in turn, but here the inspection of three networks will be used as examples of information that could be gained by a medicinal chemist.

In one such network, shown in Figure 4, several suggested compounds are shown resulting from matched series comprised of different combinations of 5 input compounds; the matched series of compounds forming the basis for one of these suggestions is highlighted. In this case, the links between all 5 compounds are highlighted, but for some of the suggestions only 3 compounds are used as only those matching derivatives are found in other chemotypes. The details for the suggestion are shown in Figure 5, where the R-groups that make up the matched series are displayed in ascending order of activity (right to left) for the two chemotypes involved, the 5 derivatives in the naphthyl series as the top row and the 6 derivatives of the benzothiophene series as the second. The final benzothiophene cyclopropylmethylether derivative is the most active in this SAR series and is missing from the naphthyl series; the perfect correlation in the activity order of the R-groups gives confidence that the cyclopropylmethylether derivative will give an improvement in activity in the naphthyl series.

In this case, a high correlation between these two series would be expected as the thiophene ring has been used as a phenyl surrogate on many occasions [12]. However, it is the easy identification of these missed opportunities that makes this analysis method valuable.

(a)



(b)

Figure 3. (a) The MSA output from the analysis of the PPARγ data set, where each matched series from the input data is shown as a vertical line of white cards, each representing an input compound; the new compound suggestions (based on the series within the data set that correlate with this matched series) are shown as coloured cards. Different colours represent different series from which the suggestion was derived. If a network contains more than one vertical line of white cards this indicates that several matched series have been identified from subsets of the displayed white cards and the new compound suggestions derived from these different subset series are given different colours within the network. (b) The MSA output where the card representing each new compound suggestion is coloured by the maximum correlation of the activity in the matched series to the other series found in the input data set. The colour range is red = low correlation (0.2) to yellow = high correlation (1.0).

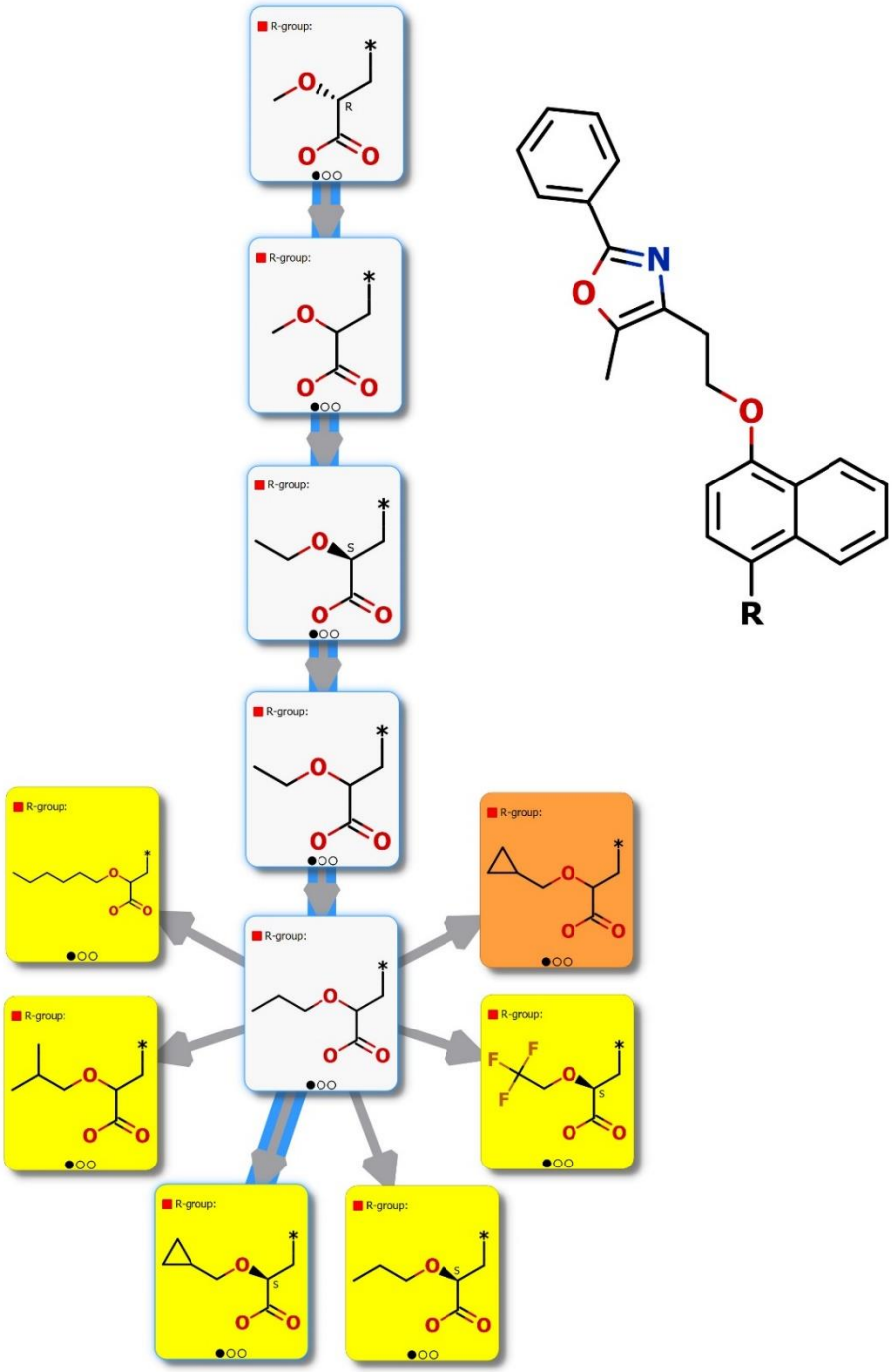
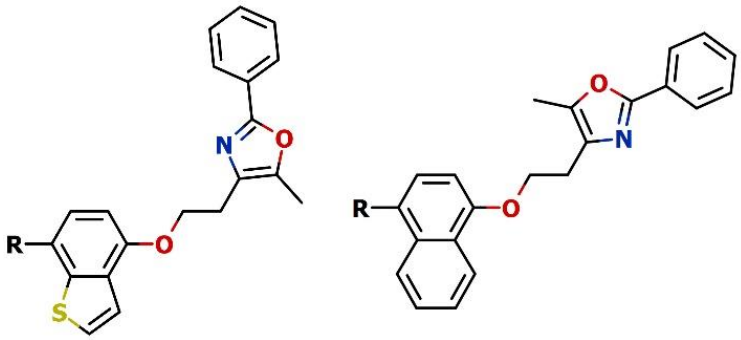


Figure 4 A close-up view of one of the networks from the MSA of the PPARγ data set, where each compound is represented by a card on which the R-group that varies in the identification of the series is shown. The common scaffold for the series is shown inset. The novel, suggested compounds are coloured by the correlation with the matched series that inspired them, from red = low correlation (0.2) to yellow = high correlation (1.0), and the white cards are the compounds which made up the matched series. By selecting one of the suggestion cards (in this case the cyclopropylmethyl ether) the links between the white cards, which form the matched series leading to the suggestion, are highlighted.



(a)

Scaffold	Correlation						
			8.155	7.523	7.444	6.947	5.804
	1	8.699	8.398	8.155	7.678	7.337	5.842

(b)

Figure 5 (a) The benzothiophene and naphthalene scaffolds indicating the point of substitution for the matched series. (b) An example of perfect SAR transfer between two series, the benzothiophene on the bottom row and the naphthalene series on the top row. The suggestion of a cyclopropylmethyl ether derivative in the naphthalene series is based upon the correlated activity of the 5 derivatives in the naphthalene and benzothiophene series shown in the table and the improved activity of the cyclopropylmethylether derivative in the benzothiophene series.

2) Overlay hypotheses

The underlying premise of MSA is that if the activity order seen for a set of derivatives in one chemotype matches the same series in a different chemotype then the varying R-groups within the series are binding in highly similar environments. The MSA performed on the set of CDK2 compounds was chosen in order to illustrate how SAR transfer could help validate existing, or even propose new, binding hypotheses and overlays between different chemotypes. There are many published structures of this target with co-crystallised derivatives in the RCSB PDB [13] that enables some validation of this approach [14]. Therefore, in this example, we will focus on sub-networks resulting from the MSA that contain compounds with published co-crystal structures and the diagram in Figure 6 illustrates one such network.

Figure 6 has a grey card in the network to indicate that the chloro derivative has been co-crystalised with CDK2 and the yellow cards are the new suggestions based upon the MSA. In Figure 7(b) the matched series of diamino-pyrimidine derivatives, and those from a 2-indolone series (shown in Figure 7(a)), with which they correlate, are listed and the activity increase in a 2-indolone series suggests that derivatisation of the aniline aryl in the diamino-pyrimidine series would be beneficial. An overlay of the aniline aryl ring with the indolone fused aryl is possible but would have to accommodate the meta versus para relationship of the substituents to the aniline nitrogen from the two matched series. A definitive overlay between these two series is non-obvious (see Supporting Information) and without the supporting evidence from the 2-indolone series it would be unlikely that the SAR series of H, Methyl, Chloro, in the diamino-pyrimidine series would lead to 4-pyridylmethyl,amido as a possible next derivative. However, as can be seen from

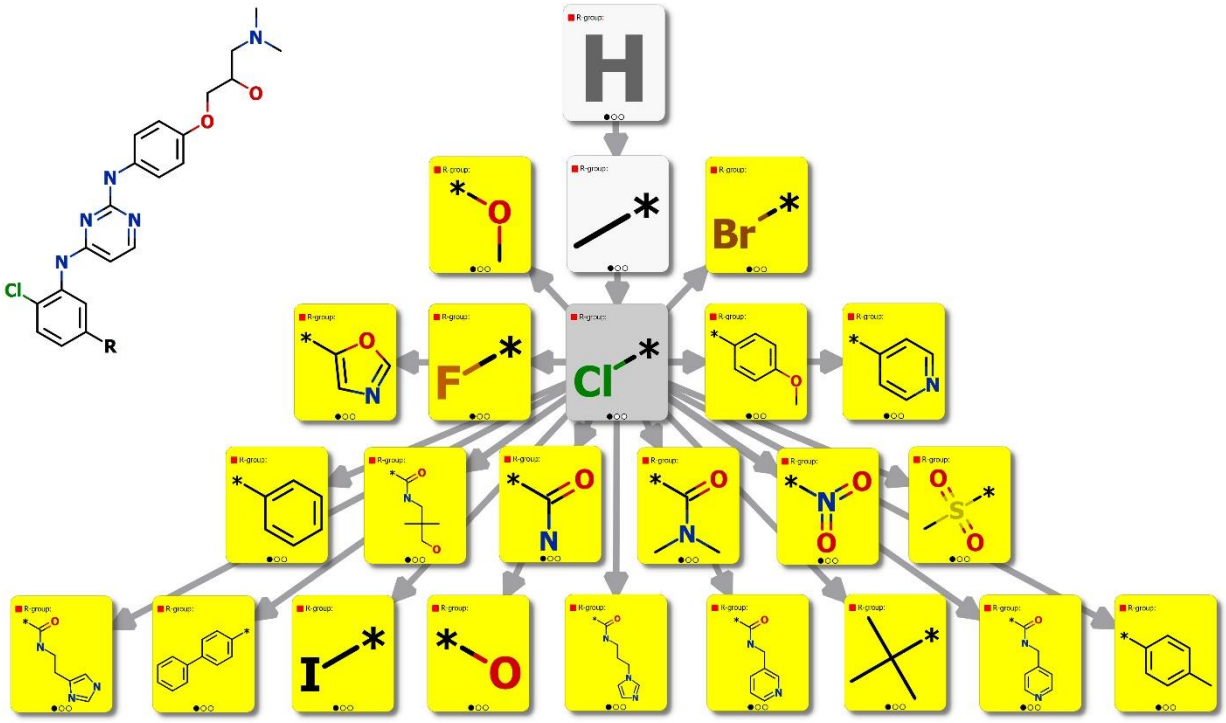
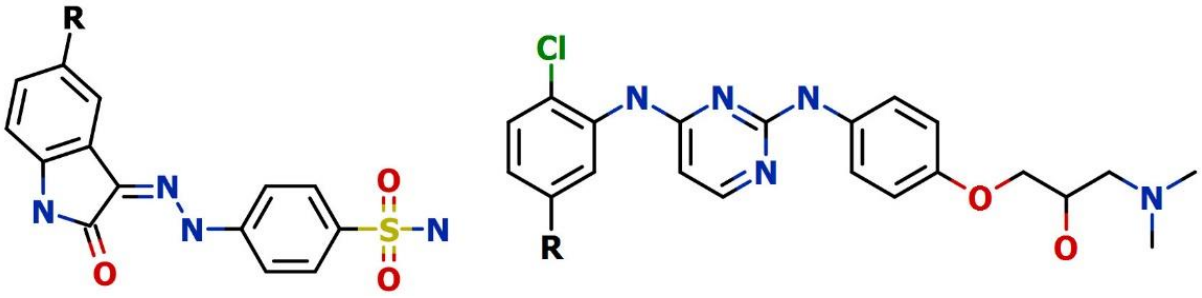


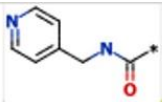
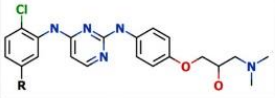
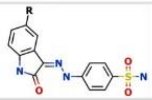
Figure 6 A sub-network with an extensive set of suggestions that was created from a MSA of the large CDK2 data set. Each compound is represented by a card on which the R-group that varies in the identification of the series is shown and the common scaffold for the series is shown inset. The novel, suggested compounds are coloured yellow, reflecting a perfect correlation with the matched series that inspired them; the white and grey cards are the compounds which made up the matched series. The grey coloured card is the chloro derivative from the diamino-pyrimidine series that has been co-crystallised with the Human CDK2 protein and the coordinates deposited in the RCSDb (PDB ID: 1H01 [16])

the crystal structures of the two scaffolds used in this case (see Figure 7(c)), the overlay implied by this SAR crossover is confirmed (albeit retrospectively). From this overlay it is clear that a larger substituent would force a rotation about the aniline bond and direct the meta substituent into the same area as that occupied by the pyridylmethyl,amide off the 5-position of the 2-indolone, but the suggested derivative would have a reasonable chance of being the most active in the diamino-pyrimidine series. Of course, it would be necessary to test this hypothesis through the synthesis of additional derivatives in the series.

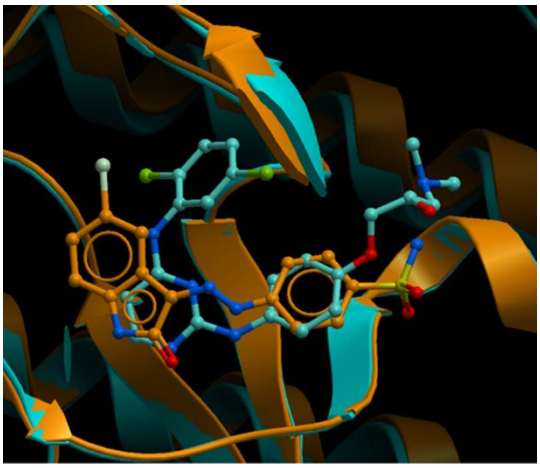
A similar scenario can be found within the MSA of the PPARγ data set. In Figure 8, the selected matched series of three azetidinone N-aryl derivatives correlate perfectly with the aryl derivatives of a glycine-carbamate series. If one examines the details of this matched series (Figure 9) one sees that the aryl derivatives in these series are at different ends of the molecules. If one assumes a particular overlay of these chemotypes that follows a crude 2-point pharmacophore of an acid interaction separated from a hydrophobic interaction by a spacer, then this apparent SAR crossover may appear irrelevant. However, this is still an assumption and this analysis could stimulate other hypotheses, analogous to the CDK2 example, such as that the two aryl rings are binding in the same part of the protein. This would mean that the binding mode would have to change when the neighbourhood of the carboxylic acid changes, such that the hydrophobic end of the glycine-carbamate series could bind in the same space as the azetidinone N-aryl substituent. The flexibility of the linker therefore might be necessary to allow this and hence rigidification of the linker might change the SAR around this aryl ring in the glycine-carbamate chemotype.



(a)

Scaffold	Correlation		Cl*		H
			6	5.301	5.222
	1	8.051	7.523	7.337	6.921

(b)



(c)

Figure 7 (a) The 2-indolone and diaminopyrimidine scaffolds indicating the point of substitution for the matched series. (b) An example of perfect SAR transfer between two series, the 2-indolone on the bottom row and the diaminopyrimidine series on the top row. The suggestion of a pyridylmethyl,amido derivative in the diaminopyrimidine series is based upon the correlated activity of the 2-indolone derivatives shown in the table. The use of the much larger sidechain in the diaminopyrimidine series would be non-obvious, based upon the SAR series of smaller substituents, without the support given by the 2-indolone series. (c) An overlay of the crystal structures with PDB codes 1H01 (cyan ribbon and carbon atoms) and 1FVT (orange ribbon and carbon atoms) showing the inhibitor binding site and the diaminopyrimidine and 2-indolone scaffolds shown in (a). The crystal structure for the dichloroanilino,diaminopyrimidine derivative has the meta chloro group directed towards and contacting the Glycine rich loop region however a larger substituent would force a rotation about the aniline bond and direct the meta substituent into the same area as that occupied by the pyridylmethyl,amide off the 5-position of the 2-indolone (indicated by the bromo derivative found in the 1FVT structure [17,18]).

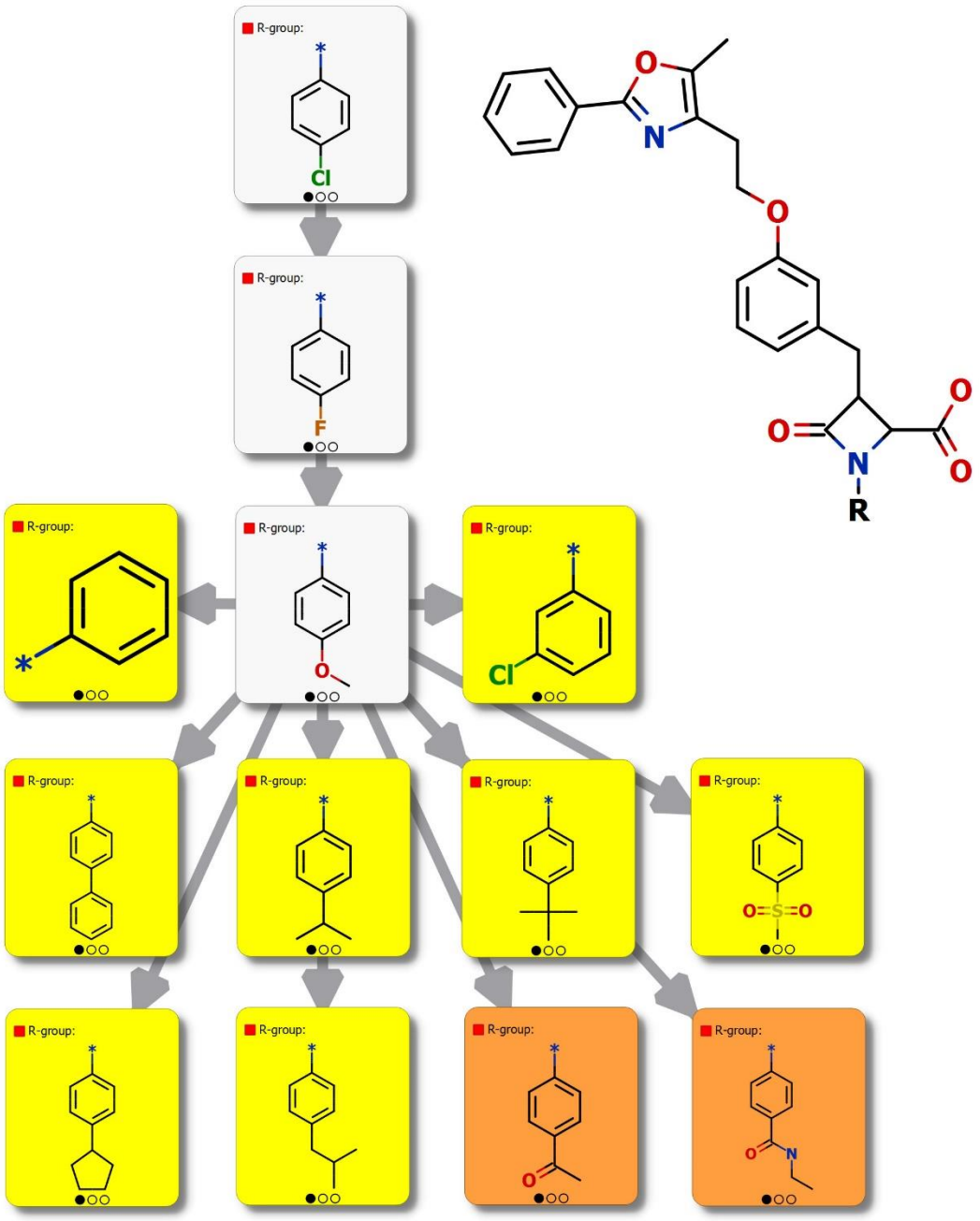
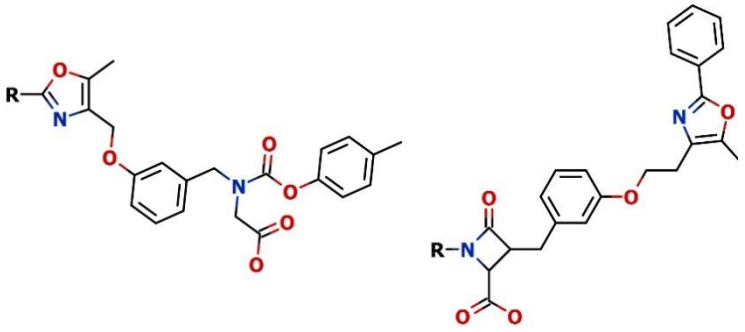
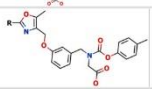


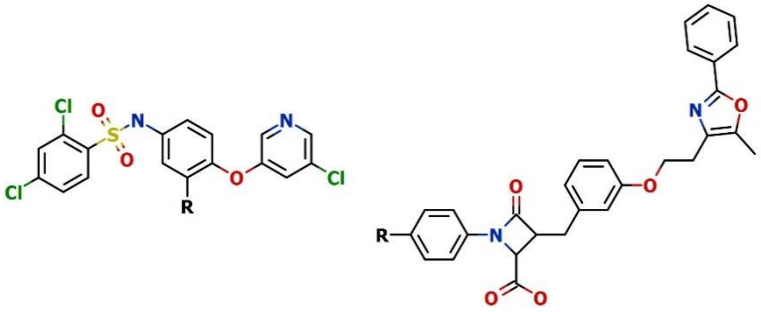
Figure 8 A close up view of another of the networks from the MSA of the PPARγ data set that details a series based around an azetidinone carboxylate. Each compound is represented by a card on which the R-group that varies in the identification of the series is shown and the common scaffold for the series is shown inset. The novel, suggested compounds are coloured by the correlation with the matched series that inspired them, from red = low correlation (0.2) to yellow = high correlation (1.0), and the white cards are the compounds which made up the matched series.

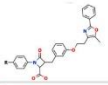
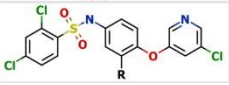


(a)

Scaffold	Correlation				
			6.796	6.745	6.444
	1	6.304	6.059	5.638	5.555

(b)



Scaffold	Correlation				
			6.796	6.745	6.444
	0.5	7.824	7.721	7.444	7.678

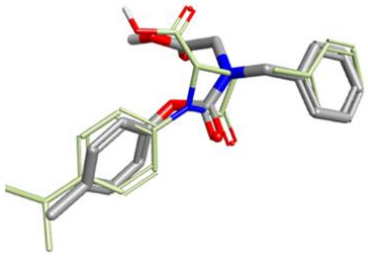
(c)

Figure 9 (a) The glycine-carbamate and azetidinone scaffolds indicating the point of substitution for the matched series. (b) An example of perfect SAR transfer between two series, the azetidinone series (on the top row) and the glycine-carbamate series (on the bottom row). The suggestion of an isopropyl derivative in the azetidinone series is based upon the correlated activity of the three derivatives shown in the table. (c) The diaryl sulfonamide and azetidinone scaffolds indicating the point of substitution for the matched series. However, this suggestion is based upon a poor correlation between the activities and given the lack of a reliable overlay (as the diaryl sulphonamides are unlikely to be acidic) this would lower the priority of this suggestion.

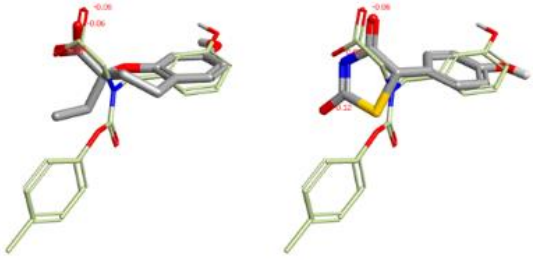
This use of the SAR transfer methodology to challenge the project assumptions can be a very useful mechanism to open up new chemotypes or provide fresh thinking within a project.

However, given the short matched series and the small differences in the measured values, which almost certainly lie within the variability of the assay, the most likely explanation of this example is one of coincidence. As stated earlier, the basis of MSA is that the SAR of substituents will correlate if they are in the same binding environment. This usually means that they are binding in the same binding pocket of a protein (as shown by the CDK2 example), but in this case, it is likely that the protein surrounding the two aryl regions at each end of the molecule provides two very similar binding environments, which are distinct in space but not in character. Therefore, it is likely that the azetidinone N-aryl ring occupies the same binding region as the carbamate aryl ring of the glycine-carbamate series (as demonstrated in Figure 10(a)) and the activity order at the oxazole aryl ring happens to match that of the azetidinone N-aryl.

The above example is in concordance with the underlying assumption embodied by the Topliss tree approach to drug design, as mentioned in O'Boyle *et al.* [3]. One can infer the nature of the binding pocket from a particular activity order for the substituents that are bound within. Therefore, by knowing what has worked in the past the Topliss tree can make suggestions as to what substituents should be an improvement. The network shown in Figure 8 is effectively an automatic generation of a Topliss tree-like analysis, building on the specific, project-related experience. It is also of interest to note that other chemotypes within this data set have correlating matched series within the network shown in Figure 8 and these thiazolidinedione or benzohydropyran scaffolds do not have a substituent in the aryl binding area (exemplified in Figure 10(b)). Hence, this may be an area which these chemotypes could explore to either improve affinity, or allow the removal of other substituents that might be causing issues in these chemotypes.



(a)



(b)

Figure 10 (a) The most likely overlay of the glycine-carbamate (shown in grey capped sticks) and azetidinone (shown in thin green sticks) scaffolds showing the correspondence between the carbamate aryl and the N-aryl substituent from the azetidinone. (b) Similar overlays with the thiazolidinedione and benzohydropyran series (shown in grey sticks) with the glycine-carbamate scaffold (shown in thin green sticks) highlighting the opportunity to expand into the area occupied by the aryl carbamate[19].

Finally, within Figure 8 one can see that there are two orange coloured cards that correspond to poor correlation values and these suggestions are derived from a correlation with a set of derivatives in a diaryl sulfonamide series (detailed in Figure 9(c)). Based on the crude pharmacophore mentioned above, the diaryl sulfonamides have the hydrophobic groups but do not have an obviously acidic centre (the sulfonamide NH would become acidic as the electron withdrawl of the aryl rings increased but with simple phenyl substituents, the expected pKa would be >7), hence the low correlation and the lack of an obvious overlay to the simple pharmacophore would make these suggestions of lower priority.

3) Erroneous assay point detection

Figure 11 shows a network from the PPARγ set which indicates that the SAR correspondence between series switches from good, to poor, and back to reasonable again, depending on the length of the SAR series being considered. The top row of yellow cards is based on a series of only 3 derivatives and matches changes in the hydrophobic tail region of the benzohydropyran series with that of the glycine carbamate derivative series as seen in Figure 12(a).

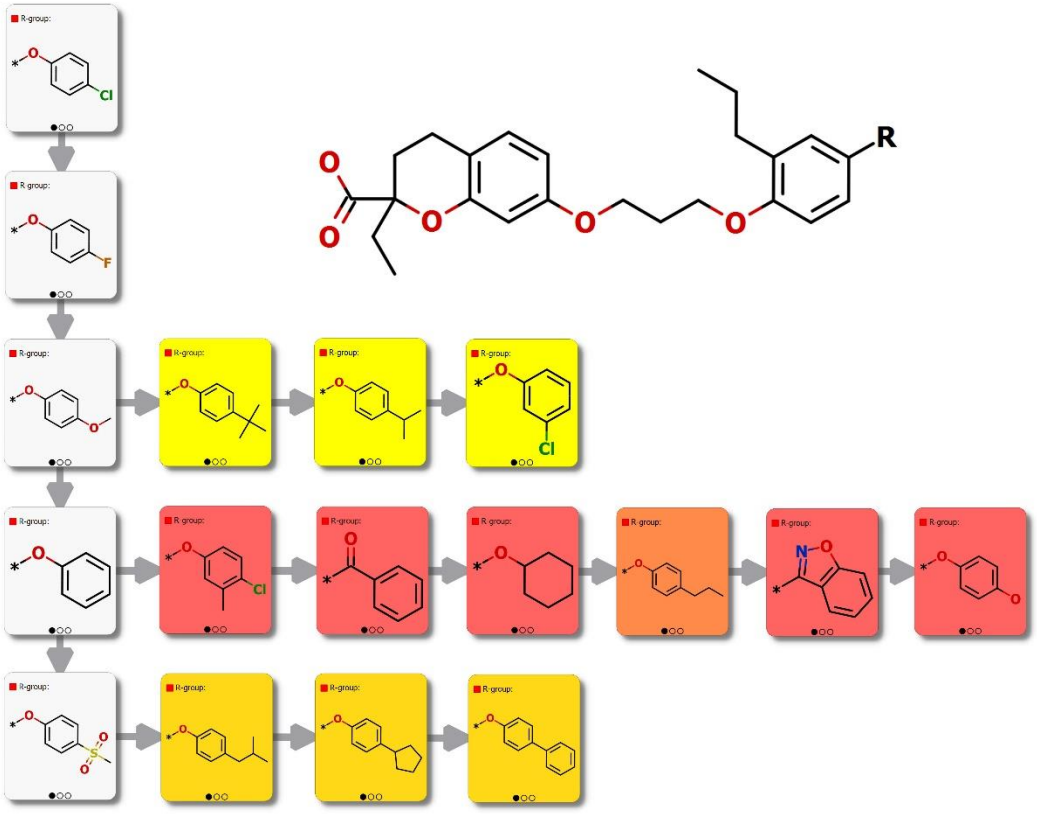


Figure 11 A close-up view of another of the networks from the MSA of the PPARγ data set that details a matched series based around a benzohydropyran scaffold. Each compound is represented by a card on which the R-group that varies in the identification of the series is shown and the common scaffold for the series is shown inset. The novel, suggested compounds are coloured by the correlation with the matched series that inspired them, from red = low correlation (0.2) to yellow = high correlation (1.0), and the white cards are the compounds which made up the matched series.

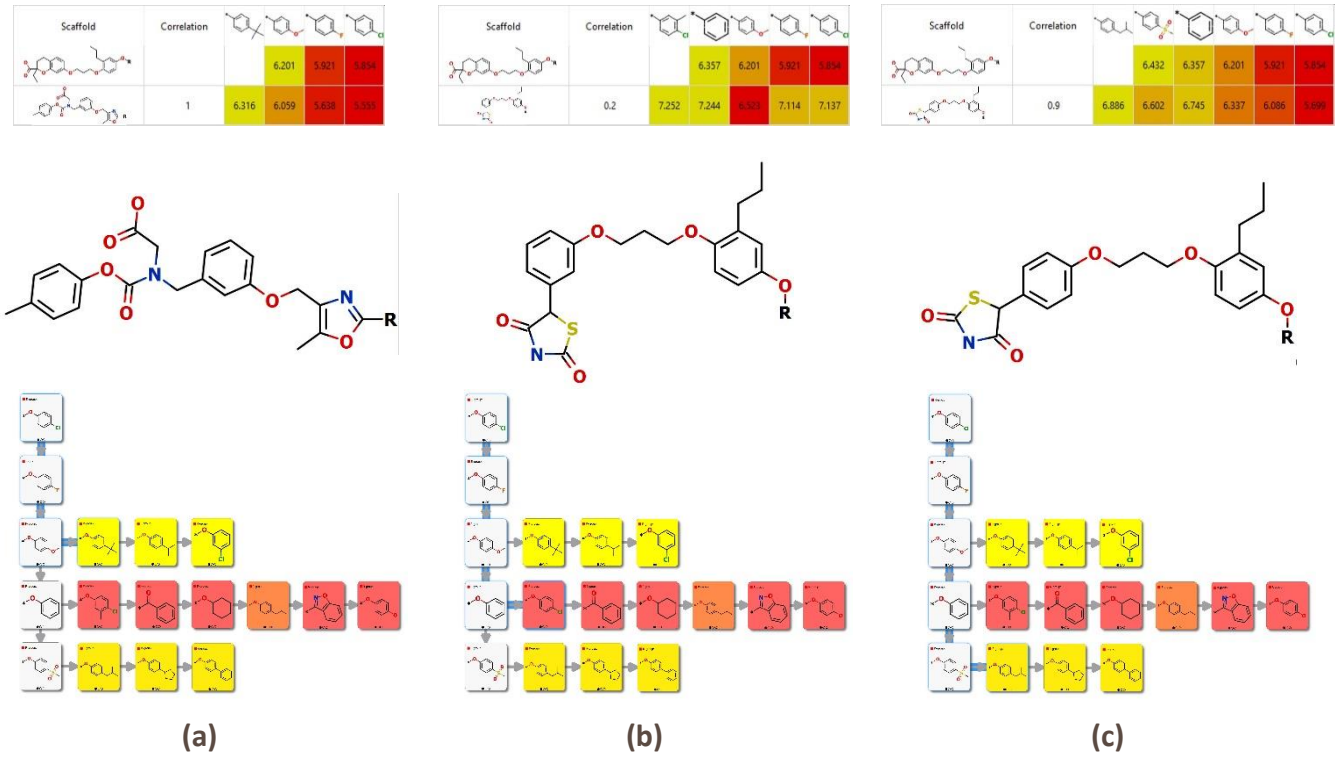


Figure 12 MSA output from the exploration of the network shown in Figure 11. For each example the upper panel shows the correlation of activities between the matched series giving rise to the suggestion and the correlated series. Below this, the scaffold for the correlated series is shown enlarged for convenience, which the lowest panel shows the highlighted network to show the path taken for each suggestion, corresponding to the different matched series on the benzohydropyran scaffold. From this it can be seen that these activities correlate in (a) and (c), but therethere is a lack of correlation in (b) due to the presence of an obvious outlier in the lower row. (a) The t-butylphenyl derivative was found in the glycine-carbamate series. (b) The 4-chloro,3-methylphenyl derivative was found in the meta- thiazolidinedione series once the phenyl derivative in the benzohydropyran series is included, although the methoxyphenyl derivative in the meta- thiazolidinedione series lowers the correlation dramatically. (c) The sec-butyl,phenyl derivative was found in the para- thiazolidinedione series given the extension of the benzohydropyran series with the 4-methylphenylsulphone derivative. The methoxyphenyl derivative in this series is now more in line with the activities of the other derivatives.

However, when the fourth derivative is included in the benzohydropyran series the SAR suddenly does not match, because the corresponding series has now changed to the meta thiazolidinedione series (as seen in Figure 12(b)). From the inspection of one of these thiazolidinedione derivatives one can see that the central anisole derivative has much lower activity than one would expect. Upon extending the benzohydropyran series one further, the correlation returns to a reasonable level because, again the corresponding series has changed, this time to the para thiazolidinedione series (Figure 12(c)).

It should be noted that the activity differences between compounds in these SAR series are small and possibly within the error of an assay, so lower correlations could occur just by chance; however, the reduction in activity of the methoxy derivative in the meta thiazolidinedione series dramatically reduces the correlation and seems anomalous. Given the good correlations seen between the benzohydropyran series and the other two series, this suggests that the inhibition measurement for the para-methoxy aryl derivative in the meta thiazolidinedione series (and perhaps the other values in this series as their spread is much narrower than in the other series) should be repeated.

Conclusion

We have shown that the analysis of the correlating and non-correlating matched series, using the SAR transfer methodology, enables detailed hypotheses regarding compound alignments or their binding modes to be formulated and highlights anomalous data points in a way that is complementary to simple MMPA. The advantage of using this methodology across the project's own data is that the suggestions generated are more likely to be considered relevant to the project. The project chemists can make better decisions regarding the relationships between their different chemotypes and monitor those decisions as more compounds are made and tested. As projects enter the later stages of optimisation, this analysis will enable the selection of the most useful compounds to synthesise, to improve activity, test a binding hypothesis or if a leading chemotype unexpectedly become unsuitable for further development. The network layout of the results as cards containing structural information [11] also makes visualisation of the results a facile process and large datasets (several thousands of compounds), typical of the size generated within a long running project, can be analysed very quickly with the Matsy/SAR transfer methodology.

The numbers of suggestions for new compounds generated by MSA may be too large for easy consideration by a medicinal chemist; the examples herein generated hundreds or thousands of suggestions. Therefore, to reduce the numbers to a more manageable level, the suggestions can be prioritised by the weight of evidence for each suggestion, e.g. the correlation with a previously observed matched series or the number of observations of the same matched series. Furthermore, the suggestions are based only on improvement in potency, therefore it may be useful to predict other factors such as physicochemical and ADME properties and prioritise the ideas against a multi-parameter profile of properties required for a high-quality compound [15].

Future Perspective

The generalisation of MMPA to the longer series of MSA offers the ability to make more relevant predictions of new substitutions that are likely to improve target activities. Furthermore, as illustrated herein, analysis of correlations between matched series can reveal new binding hypotheses and strategies for optimisation, as well as provide a method for checking the consistency of experimental data. We expect to see the popularity of MMPA to translate into a similar uptake of MSA as software becomes more widely available to provide access in an accessible and intuitive way.

To date, MSA has been applied primarily to the analysis of target activity data. However, the method applies equally well to other properties, such as absorption, distribution, metabolism and elimination (ADME). The challenge here has been the availability of sufficiently large, SAR rich data sets on which to build a database of matched series, similar to that used herein based on ChEMBL activity data. Such databases are available within large pharma companies, but access to similar sets in the public domain remains limited. Hopefully, an increase in acceptance of pre-competitive data sharing will lead to the availability of appropriate databases in the future. Databases of matched series provide a particular advantage for this, in that they do not require the disclosure of full structures, only the changes in substituent and commensurate changes in the measured property.

Finally, the analysis of 2-dimensional (2D) SAR by analysis of experimental data, using methods such as MSA can reveal interesting correlations and trends relating compound structure and activity. However, these are best understood in the context of the 3-dimensional (3D) binding of the ligand to the protein target, as illustrated in the second example above. Linking 2D and 3D SAR in a highly visual and interactive way will enable more rapid interpretation and exploitation, leading to quicker compound optimisation.

Supporting Information

PPARγ and CDK2 data sets used in this analysis in csv text file format. Torch3D based overlays of the CDK2 cocrystallised ligands. Torch3D based overlay methodology of the head groups from the PPARγ data set shown in Figure 10.

References

(1) Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, McFaul P, Wood JM, Coclough N, Law B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med. Chem.* 49(23), 6672-6682 (2006)

(2) Wawer M, Bajorath J. Local Structural Changes, Global Data Views: Graphical Substructure − Activity Relationship Trailing. *J. Med. Chem.* 54(8), 2944−2951 (2011)

(3) O'Boyle NM, Bostrom J, Sayle RA, Gill A. Using Matched Molecular Series as a Predictive Tool to Optimize Biological Activity. *J. Med. Chem.* 57(6), 2704-2713 (2014)

(4) Peltason L, Weskamp N, Teckentrup A, Bajorath J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* 52(10), 3212-3224 (2009)

(5) Hussain J, Rea C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* 50(3), 339-348 (2010)

(6) Gupta-Ostermann D, Wawer M, Wassermann AM, Bajorath J. Graph mining for SAR transfer series. *J. Chem. Inf. Model.* 52(4), 935-942 (2012)

(7) Zhang B, Wassermann AM, Vogt M, Bajorath J. Systematic assessment of compound series with SAR transfer potential. *J. Chem. Inf. Model.* 52(12), 3138-3143 (2012)

(8) StarDrop, Optibrium Ltd, Cambridge UK, http://www.optibrium.com/stardrop/

(9) Papadatos G, Overington JP. The ChEMBL database: a taster for medicinal chemists. *Future Med Chem.* 6(4), 361–364 (2014)

(10) Whilst it is theoretically possible to search for anti-correlated series by specifying a negative correlation value the likelihood of such series being contained in this small data set is presumed to be tiny given the fact that all these compounds have activity measured at the same target.

(11) Segall M, Champness E, Leeding C, Chisholm J, Hunt P, Elliott A, Garcia-Martinez H, Foster N, Dowling S. Breaking free from chemical spreadsheets. *Drug Discov. Today* 20(9), 1093-1103 (2015)

(12) Maxwell NA. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. *J. Med. Chem.* 54(8) 2529-2591 (2011)

(13) Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE The Protein Data Bank. *Nucleic Acids Res.* 28(1), 235-242 (2000)

(14) Posy SL, Claus BL, Pokross ME, Johnson SR. 3D Matched Pairs: Integrating Ligand- and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* 53(7), 1576–1588 (2013)

(15) Segall MD. Multi-Parameter Optimization: Identifying high quality compounds with a balance of properties. *Curr. Pharm. Des.* 18(9), 1292-1310 (2012)

(16) PDB ID: 1H01 - Beattie JF, Breault GA, Ellston RPA, Green S, Jewsbury PJ, Midgley CJ, Naven RT, Minshull CA, Pauptit RA, Tucker JA, Pease, J. E. Cyclin-Dependent Kinase 4 Inhibitors as a Treatment for Cancer. Part 1: Identification and Optimisation of Substituted 4,6-Bis Anilino Pyrimidines. *Bioorg. Med. Chem. Lett.* 13(18), 2955-2960 (2003)

(17) PDB ID : 1FVT - Davis ST, Benson BG, Bramson HN, Chapman DE, Dickerson SH, Dold KM, Eberwein DJ, Edelstein M, Frye SV, Gampe Jr RT, Griffin RJ, Harris PA, Hassell AM, Holmes WD, Hunter RN, Knick VB, Lackey K, Lovejoy B, Luzzio MJ, Murray D, Parker P, Rocque WJ, Shewchuk L, Veal JM, Walker DH, Kuyper LF. Prevention of chemotherapy-induced alopecia in rats by CDK inhibitors. *Science* 291(5501), 134-137 (2001)

(18) Davis ST, Benson BG, Bramson HN, Chapman DE, Dickerson SH, Dold KM, Eberwein DJ, Edelstein M, Frye SV, Gampe Jr RT, Griffin RJ, Harris PA, Hassell AM, Holmes WD, Hunter RN, Knick VB, Lackey K, Lovejoy B, Luzzio MJ, Murray D, Parker P, Rocque WJ, Shewchuk L, Veal JM, Walker DH, Kuyper LF. Retraction. *Science* 298(5602) 2327 (2002). Note that this retraction of reference [16] relates to the failure to reproduce the biological activity of the compound in a neonatal rat model of chemotherapy-induced alopecia; therefore it does not impact the crystal structure used in the analysis herein.

(19) The alignments were generated in StarDrop [10] using the torch3D module, which is an implementation of the field based alignment methodology from Cresset http://www.cresset-group.com, Cheeseright T, Mackey M, Rose S, Vinter A. Molecular field extrema as descriptors of biological activity: definition and validation. *J. Chem. Inf. Model.* 46(2), 665–676 (2006)