# When two are not enough: Lead optimization beyond Matched Pairs

Noel M. O'Boyle,[a] Roger A. Sayle,[a] Matt Segall[b]

[a] NextMove Software Ltd., Innovation Centre, Cambridge Science Park, Milton Road, Cambridge CB4 0EY, UK

[b] Optibrium Ltd., 7221 Cambridge Research Park, Beach Drive, Cambridge, CB6 3WJ

Corresponding author: O'Boyle, N.M. (noel@nextmovesoftware.com)

## Abstract

Lead optimization projects progress by making successive enhancements to one or more starting structures. This is a classic multi-objective optimization procedure where the goal is not only to improve potency but also to improve physicochemical and absorption, distribution, metabolism and elimination (ADME) properties. For physicochemical and ADME properties, the popular matched molecular pair analysis method has been a successful strategy; however, it notably fails in the goal of improving potency. Here we discuss a lead optimization approach involving matched *series*, the extension of matched pairs to more than two R-groups, which can successfully be used to guide molecular design towards improved potency. Furthermore, this approach retains the attractive features of matched pair analysis in that it is entirely driven by experimental data and is a natural fit to the medicinal chemistry approach of designing analogs by successive small changes to an existing molecule.

## Introduction

What molecule should I make next? This is the question that occurs again and again at each step of a lead optimization project. Answering this question well may mean the difference between project success and failure, or at least between rapid progress and wasting time following numerous dead-ends.

How one decides which molecule to synthesize will clearly vary from one person to the next, but ultimately it boils down to one of two things. The first of these is the medicinal chemist's experience from working on related projects; for example, what worked last time? However, for the most part deciding what compound to make next is based on observed activity trends, from which a particular structure-activity relationship is inferred and then extrapolated to a new structure. This is commonly referred to as "chemical intuition", but in fact relies on a chemist's knowledge of potential structure-activity relationships and the relative property values of common R-groups.

Matsy [1] is a lead optimization strategy that combines both of these approaches and, rather than relying on an individual or group's experience, it uses the experience garnered by tens of thousands of medicinal chemists and available in the literature. Instead of inferring structure-activity relationships using "intuition", it bases them on this broader experience so that all predictions are made on the basis of previously observed experimental results.

## Matched Pairs and Series

The starting point for this method is the concept of a matched pair (or more formally, a Matched Molecular Pair or MMP) although, as we shall see, such MMPs are not in themselves sufficient for this purpose. A MMP refers to two molecules with the same scaffold but different R-groups at the same position [2], and has become very popular in recent years for rationalising trends in SAR [3,4]. The success of this approach is due to the fact that relative changes in property values are easier to predict than absolute values. It also fits very well with the common lead optimization procedure of changing R-groups, while keeping the underlying scaffold constant.

Predictions based on this approach work well for physicochemical properties as well as for biological activities that correlate highly with such properties. However, in general, MMP analysis does not work well for predicting R-groups that improve biological activity. This was most clearly shown in a 2008 study by Hajduk and Sauer at Abbott [5] for MMP data drawn from a broad range of targets. Potency changes associated with most MMP transformations were found to be nearly normally distributed around zero. The simple reason for this limitation of MMP analysis is that for one binding site environment changing group A to group B may increase activity, while for another binding site environment it may decrease activity. While attempts have been made to address this problem, for example by focusing on MMPs from just the target of interest [6] or with a particular atom environment [7], the underlying problem remains.

But all is not lost. If we revisit the analysis by Hajduk and Sauer, we can show that there may be a way forward. Let us take as an example those assays in the ChEMBL database [8] (https://www.ebi.ac.uk/chembl/) with $pIC_{50}$ activity data for compounds with ethyl, propyl and butyl as substituents at the same location on a scaffold. Figure 1a shows the $pIC_{50}$ for the ethyl analog versus that of the butyl, and sure enough we see a symmetric distribution of the activities around zero. In other words, changing ethyl to butyl is equally likely to increase activity as to decrease activity, consistent with the results found by Abbott.

However, what if we include additional knowledge about the context of the matched pair transformation? For example, suppose that we already know that the propyl analog has a greater $pIC_{50}$ than the butyl for our scaffold of interest. If we take the subset of the data in ChEMBL where the propyl analog is more active than the butyl, and then regenerate the original plot (Figure 1b), the distribution of the ethyl minus butyl activities is now shifted to the right away from zero. In other words, knowing that the propyl is more active than the butyl dramatically increases the chance that ethyl is also more active. More generally, if we already know additional information about activities, it should improve our ability to predict the effect of a given R-group replacement.

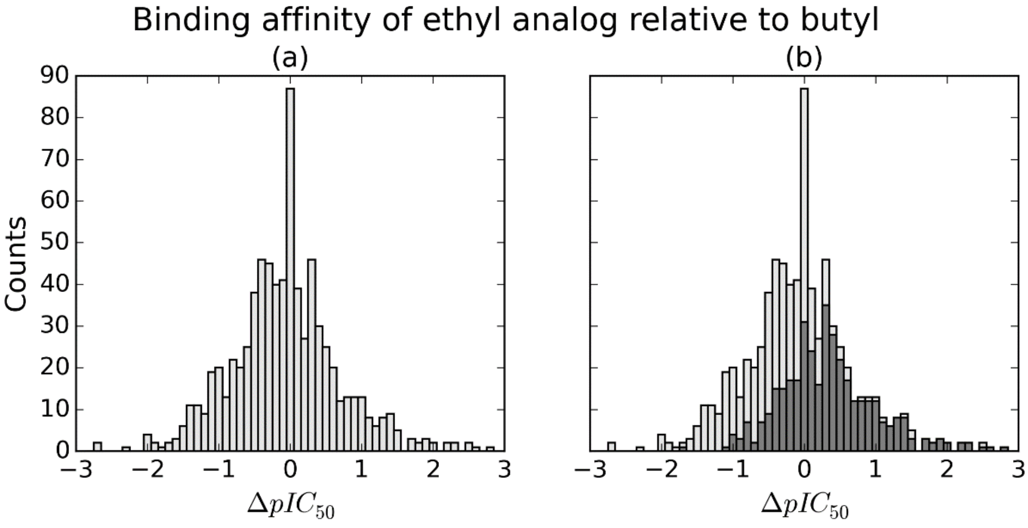


**Figure 1 – Example of improved prediction of activities given prior knowledge. (a) Histogram of data from ChEMBL 20 showing the relative binding affinity of ethyl analogs relative to the corresponding butyl analog (for scaffolds where the ethyl, propyl and butyl analog have been measured and are not inactive). (b) The additional histogram drawn in bold shows the subset of the original data where the propyl analog has greater binding affinity than the butyl.**

The question is, how best to do this? One approach would be to throw more matched pairs at the problem; rather than simply considering two R-groups and the associated MMP, for any three R-groups, consider the three associated MMPs. However, this quickly becomes unwieldy once one progresses to four R-groups and the associated six MMPs or even longer series with larger numbers of combinations.

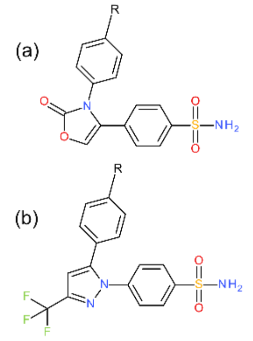
In fact, a much simpler and more elegant approach is to consider all of the associated R-groups as parts of a single Matched Molecular Series (MMS), a concept introduced by Bajorath in 2011 [9]. This is simply a generalisation of the MMP concept to a series of any length, that is, N molecules with the same scaffold but different R-groups at the same position. With a matched pair, we are asking the question "Will changing B to C increase the activity?"; in contrast, if using a matched series of length 3, we are asking "Will changing B to C increase the activity, given that B is more active than A?" In other words, using longer series introduces a context regarding a particular binding site environment.

Although the term matched pairs was first described in 2005 [2], and the limitations of the approach were already shown by Hajduk and Sauer in 2008, it is interesting to ask why it took so long to start looking beyond pairs to longer series? One hypothesis is that by naming the concept using the term "pair", chemists focused on thinking in terms of two R-groups exactly and found it difficult to think outside this box. Furthermore, the concept of matched pairs has become synonymous for many with "a matched pair transformation" (that is, a replacement of a terminal R Group), and this cemented the idea of two R-groups as a fundamental concept, rather than just a specific instance of a general case.

The following sections describe two approaches to guide lead optimisation using MMS, namely SAR Transfer and Matsy. In both cases, it will be apparent that such predictions are at their least reliable when based on matched pair data rather than data from longer matched series.

## SAR Transfer predictions

The concept of SAR Transfer, as introduced by Bajorath [10], is best explained with an example. Suppose that we have synthesised the set of 8 analogs shown in Figure 2 that have different R-groups at a particular location on a common scaffold (a). This, of course, is a MMS of length 8. Having measured the biological activities of these analogs, we need to decide what R-group to make next. One approach to do this is to search a database of biological activities to find MMS containing the same 8 R-groups (or a large subset thereof) and where the order of the activities of the analogs is a close match to the original series (this can be measured using rank correlation). Having found such a match, e.g. the scaffold (b) and its corresponding R-groups, it is a reasonable assumption that any additional R-groups in this new series that have improved activity relative to the original 8 R-groups, may also further improve the activity for the original series involving scaffold (a). In other words, we are transferring SAR from a database match to our own series. For the particular case in the example, the $NH_2$ and SMe groups may offer improved activity for scaffold (a) based on the match to scaffold (b).



| | pIC$_{50}$ | | |
|---|---|---|---|
| **R** | (a) | | (b) |
| SMe | ?? | ← | 5.92 |
| $NH_2$ | ?? | ← | 5.88 |
| OMe | 6.68 | — | 5.59 |
| Me | 6.10 | — | 4.82 |
| Cl | 5.92 | — | 4.75 |
| F | 5.82 | — | 4.59 |
| Et | 5.81 | — | 4.54 |
| $CF_3$ | 5.70 | ╳ | <4.00 |
| H | 5.62 | | 4.26 |
| COOH | 4.23 | — | <3.60 |

**Figure 2 – Example of SAR Transfer. The arrows indicate potential SAR transfer from the matched series involving scaffold (b), CHEMBL772766 [11], to that involving scaffold (a), CHEMBL768956 [12]. The lines joining the columns indicate the correspondence between the rank orders. In this case the rank correlation as measured by Kendall's tau is 0.93 as all of the activities are in the same order apart from the $CF_3$ and H activities**

It should be clear that this approach is more likely to work as the number of R-groups in common increases, and the higher the correlation of the relative activities of the R-groups in the series. In particular, this is a useful technique to identify gaps that are worth exploring in a dense R-group matrix; for example, a scaffold with two R-group positions where many of the R1xR2 combinations have been synthesised and tested.

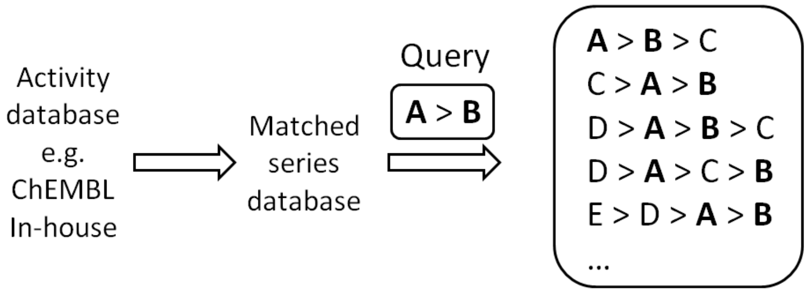
Unfortunately, the longer the MMS, the less likely it is that a match to a particular series will be found in a reference database, let alone a match with a high correlation of activities. On the other hand, if the length of the MMS is short, even if the activities have perfect correlation, you are unlikely to be confident that the SAR can be transferred to the original series from a single match in the reference database. Furthermore, the number of matches to a short series may be of the order of hundreds or even thousands. The next section describes the Matsy method, an approach that was developed to handle this situation.

## Matsy

The Matsy algorithm [1] can be considered a statistical version of the SAR Transfer method that can handle predictions based on short MMS. The origin of the method is the observation that, given a set of R-groups, certain activity orders are found more commonly in matched series composed of those R-groups. Given an existing matched series, the algorithm searches an activity database for all R-groups that have been measured along with those in the input, and calculates the percentage of times each R-group increased the activity beyond the most active R-group in the input series. The R-groups with the highest percentages are presented as the most likely candidates to try next.

Figure 3 presents this approach in the context of a MMS database where only five matches are found in the database. In this case, the R-group D had improved activity relative to the best R-group in the query (A) three times out of three, i.e. 100% of the time; in contrast, C only improved the activity once out of four times, i.e. 25% of the time. In practice a higher cut-off is applied to the number of observations so that the user can have some confidence in the results.

A more realistic example would be to search a MMS database derived from ChEMBL. Let's assume that we have synthesised a matched series in which ethyl is more active than propyl and propyl itself is more active than methyl (that is, Et > Pro > Me). The top prediction from the Matsy algorithm is cyclopentyl on the basis of 23 observations in ChEMBL of which 39% increased the activity. The next best prediction is a bromine, which increased activity 38% of 21 times. It is worth noting that swapping an ethyl with a bromine will reduce the logP; this illustrates the fact that the predictions are not solely driven by logP (a frequently asked question), but are driven by observed trends in the data.

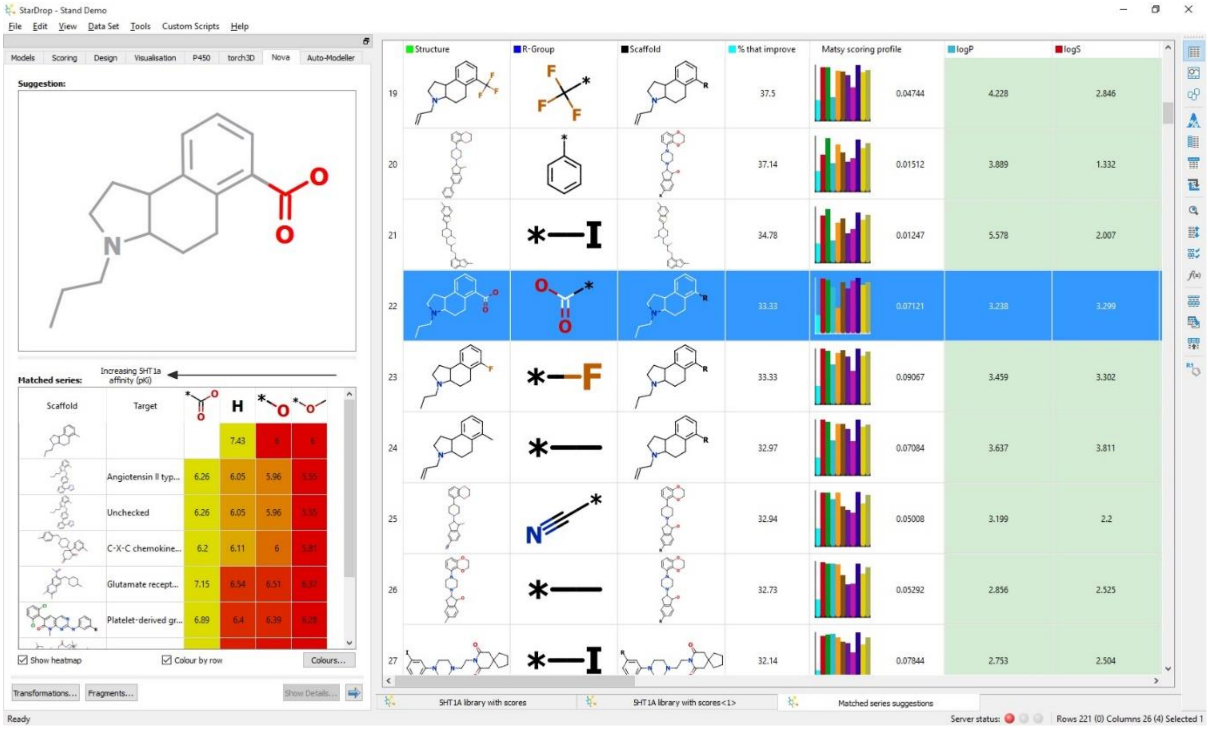


| R Group | Observations | Obs that increase activity | % that increase activity |
|---|---|---|---|
| D | 3 | 3 | 100 |
| E | 1 | 1 | 100 |
| C | 4 | 1 | 25 |
| ... | ... | | ... |

***Figure 3 – Overview of Matsy method.*** *The Matsy method works by searching a database of matched series for matches to a query matched series. In contrast to SAR transfer, only matches with perfect correlations are considered (that is, the same exact activity order). Information on R-groups in the matches is then collated into a table. This image was previously published in O'Boyle et al [1].*

## Practical application of Matched Series to guide design

As discussed, MMPs have proved to be attractive because the corresponding transformations are easily interpreted; the improved predictive power of MMS can also be accessed in an intuitive way. Existing series of compounds can be analysed to find corresponding matched series in a database and, from these, automatically generate new suggestions for optimisation. To gain confidence in the rationale for these suggestions, the underlying experimental evidence can be presented and easily explored. Coupling this with predictive modelling of other properties enables true multi-parameter optimization to quickly prioritise new compounds to pursue, as illustrated in Figure 4.



StarDrop - Stand Demo
File Edit View Data Set Tools Custom Scripts Help
Models Scoring Design Visualisation P450 torch3D Nova Auto-Modeller
Suggestion:
Increasing 5HT1a affinity (pKi)
☑ Show heatmap
☑ Colour by row
Colours...
Transformations...
Fragments...
Show Details...
Ready
5HT1A library with scores
5HT1A library with scores<1>
Matched series suggestions
Server status:
Rows 221 (0) Columns 26 (4) Selected 1

Matched series:

| Scaffold | Target | *C(=O)O | H | *O | *OC |
|---|---|---|---|---|---|
| | | | 7.43 | 6 | 6 |
| | Angiotensin II typ... | 6.26 | 6.05 | 5.96 | 5.55 |
| | Unchecked | 6.26 | 6.05 | 5.96 | 5.55 |
| | C-X-C chemokine... | 6.2 | 6.11 | 6 | 5.81 |
| | Glutamate recept... | 7.15 | 6.54 | 6.51 | 6.37 |
| | Platelet-derived gr... | 6.89 | 6.4 | 6.39 | 6.28 |

| | Structure | R-Group | Scaffold | % that improve | Matsy scoring profile | logP | logS |
|---|---|---|---|---|---|---|---|
| 19 | | | | 37.5 | 0.04744 | 4.228 | 2.846 |
| 20 | | | | 37.14 | 0.01512 | 3.889 | 1.332 |
| 21 | | *—I | | 34.78 | 0.01247 | 5.578 | 2.007 |
| 22 | | | | 33.33 | 0.07121 | 3.238 | 3.299 |
| 23 | | *—F | | 33.33 | 0.09067 | 3.459 | 3.302 |
| 24 | | *— | | 32.97 | 0.07084 | 3.637 | 3.811 |
| 25 | | | | 32.94 | 0.05008 | 3.199 | 2.2 |
| 26 | | *— | | 32.73 | 0.05292 | 2.856 | 2.525 |
| 27 | | *—I | | 32.14 | 0.07844 | 2.753 | 2.504 |

**Figure 4 – Example of the output from Matsy. The data set to the right shows the suggested compounds and the R-groups and scaffolds from which they derive, along with the percentage of observations that improve the activity. This is supplemented by predicted physicochemical and ADME properties and a score against a multi-parameter profile of property criteria. The selected compound is shown in more detail on the left, with the suggested substitution highlighted and a table below showing the experimental evidence provided by the corresponding matched series in ChEMBL.**

## Conclusion

The term Matched Molecular Pair made concrete a concept and technique that medicinal chemists had been aware of for years previously; namely that comparisons between the properties of two molecules that differ in a single substituent may be used to guide lead optimisation. However, the focus on two molecules rather than a set of molecules has hindered advances in property prediction. Now that there is an increasing awareness of Matched Molecular Series among chemists, we hope that they will start to look beyond matched pairs to matched series based techniques such as SAR Transfer and Matsy that overcome some of the limitations of matched pairs and open up new ways of thinking about, searching and predicting structure-activity relationships.

## References

1 O'Boyle, N.M. *et al.* (2014) Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* 57, 2704–2713

2 Kenny, P.W. and Sadowski, J. (2004) Structure Modification in Chemical Databases. In *Cheminformatics in Drug Discovery* (Oprea, T. I., ed), pp. 271–285, Wiley-VCH

3 Griffen, E. *et al.* (2011) Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* 54, 7739–7750

4 Dossetter, A.G. *et al.* (2013) Matched Molecular Pair Analysis in drug discovery. *Drug Discov. Today* 18, 724–731

5 Hajduk, P.J. and Sauer, D.R. (2008) Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *J. Med. Chem.* 51, 553–564

6 Gleeson, P. *et al.* (2009) ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem.* 17, 5906–5919

7 Warner, D.J. *et al.* (2010) WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* 50, 1350–1357

8 Bento, A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090

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

10 Wassermann, A.M. and Bajorath, J. (2011) A Data Mining Method To Facilitate SAR Transfer. *J. Chem. Inf. Model.* 51, 1857–1866

11 Penning, T.D. *et al.* (1997) Synthesis and Biological Evaluation of the 1,5-Diarylpyrazole Class of Cyclooxygenase-2 Inhibitors: Identification of 4-[5-(4-Methylphenyl)-3- (trifluoromethyl)-1H-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* 40, 1347–1365

12 Puig, C. *et al.* (2000) Synthesis and Biological Evaluation of 3,4-Diaryloxazolones: A New Class of Orally Active Cyclooxygenase-2 Inhibitors. *J. Med. Chem.* 43, 214–223