Lukac, I, Zarnecka, J, Griffen, EJ, Dossetter, AG, St-Gallay, SA, Enoch, SJ, Madden, JC and Leach, AG

Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs.

Turbocharging matched molecular pair analysis: optimizing the identification and analysis of pairs

Journal:	*Journal of Chemical Information and Modeling*
Manuscript ID	ci-2017-00335u.R2
Manuscript Type:	Article
Date Submitted by the Author:	13-Sep-2017
Complete List of Authors:	Lukac, Iva; Liverpool John Moores University School of Pharmacy and Biomolecular Sciences, Zarnecka, Joanna; Liverpool John Moores University School of Pharmacy and Biomolecular Sciences Griffen, Edward; MedChemica Limited, Dossetter, Alexander; MedChemica Limited, Chemistry St-Gallay, Steve; Sygnature Discovery Ltd, Enoch, Steven; Liverpool John Moores University, School of Pharmacy and Chemistry Madden, Judith; Liverpool John Moores University School of Pharmacy and Biomolecular Sciences Leach, Andrew; Liverpool John Moores University, School of Pharmacy and Biomolecular Sciences

Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs

Iva Lukac[†], Joanna Zarnecka[†], Edward J. Griffen[¶], Alexander G. Dossetter[¶], Stephen A. St-Gallay[§], Steven J. Enoch[†], Judith A. Madden[†], Andrew G. Leach[†¶]*

† School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK.

¶ Medchemica Limited, BioHub, Alderley Park, Macclesfield, SK10 4TG, UK

§ Sygnature Discovery Ltd, Bio City, Pennyfoot St, Nottingham, UK NG1 1GF

Email: a.g.leach@ljmu.ac.uk

Abstract. We have applied the two most commonly used methods for automatic matched pair identification and obtained the optimum settings and discovered that the two methods are synergistic. A turbocharging approach to matched pair analysis is advocated in which a first round (a conservative categorical approach which uses an analogy with coin flips; heads corresponding to an increase in a measured property, tails a decrease and a biased coin to a structural change which reliably causes a change in that property) provides the settings for a second round (which uses the magnitude of the change in properties). Increased chemical specificity allows reliable knowledge to be extracted from smaller sets of pairs and an assay-

specific upper limit can be placed on the number of pairs required before adequate sampling of variability has been achieved.

Introduction. Since the beginning of rational drug design, compound designers have faced the same question: "What to make next?" With the increased availability and applicability of high throughput techniques from around the mid 1990s, the amount of data and the diversity of compounds for which data exists have both increased incredibly. Despite this, the vast majority of compounds entering clinical trials fail to reach the market.[1] It is therefore imperative that drug discoverers find better ways to use the vast repository of data already available to improve their chances of success in the future.

One of the ways to decide what to make next is to analyze all of the molecules that differ only by a well-defined structural transformation and to use the effect of the structural change in the past to estimate its likely effect if applied again. This is matched molecular pair analysis, MMPA.[2] One of the hypotheses that underpin this approach is that the effect on pharmaceutical properties of a small structural change is more easily predicted than the absolute value of those properties for each molecule alone.[2,3] MMPA has been successfully applied in a range of ways: suggesting what to make next, predicting the properties of a new compound, identifying cases where a structural change has a minimal effect on key properties (like bioisosteres), or simply to increase our understanding of the links between biology and chemistry.[3] The output of MMPA can be the average size of the change in property caused by a change in structure or the proportion of cases that have seen the property change in a particular direction (which can be interpreted as similar to the probability that applying the structural change again will cause the same direction of change in activity).[4]

Recent advances in MMPA have focused on exploiting the context within which the matched pairs are found. This includes the context of the protein binding site and the context of larger sets of matched molecules (matched molecular series).[5–8] The chemical structure that is shared between a pair of molecules also provides a context for the structural change and can influence the effect of the structural change.[4,9,10] For instance, changing Ar-Me to Ar-F has a different effect on hERG blocking potency when the group that is changing is ortho to a CH_2N group (where the N is basic) compared to other contexts.[9] This suggests that sets of matched molecular pairs may be more useful when split into structurally specific sets of pairs. A key challenge for MMPA is that large sets of matched molecular pairs are required in order to achieve statistical significance but dividing these sets according to the different chemically defined contexts will reduce the size of the sets.

One way to address this dichotomy is to create much larger source datasets within which to identify matched molecular pairs. With this aim in mind, a consortium of pharmaceutical companies has recently been formed to create the SALT (Statistical Analysis of Large pharma daTasets) database.[11,12] These companies use one MMPA algorithm to analyze their corporate databases and to create a standardized encoding of the matched molecular pairs using the change in structure and the change in properties. Neither the complete structure of the molecules in the pairs nor their properties are required. These MMPA encodings are freely exchangeable because they do not contain any proprietary information. When these sets of matched molecular pairs are brought together, the effect of structural changes upon the same properties measured in different companies are merged and can be either positively or negatively cooperative. The details of this

process are described elsewhere.[12] A by-product of this merging is that a large set of data is available that can be used to address how best to perform this MMPA.

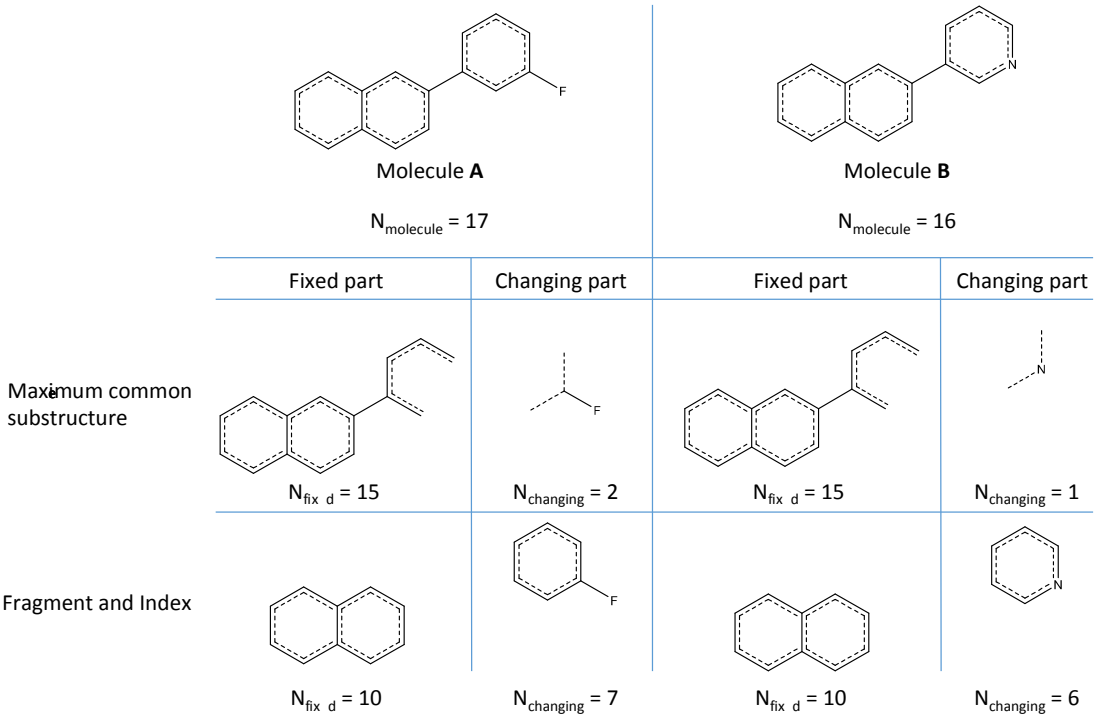

Figure 1. The pair of molecules A and B and the fixed and changing parts identified by the MCSS and F+I methods.

There are two common approaches to identifying matched molecular pairs in an automated fashion. These are the maximum common substructure (MCSS) method and the fragment and index (F+I) method. The first identifies the maximum common substructure that is shared between two molecules (Figure 1). This defines the "fixed" part while the remainder of each molecule defines the "changing" part. This is the basis of the WizePairZ algorithm.[13] The alternative F+I method identifies acyclic single bonds and cleaves them in a combinatorial fashion (often there can be very many such bonds and a limit needs to be placed on how many

are broken at once). Once each molecule is broken down into all its possible fragmentations, simple text searching can identify fragments that are shared between two molecules and this defines the "fixed" part with the remaining fragment(s) of each molecule being the "changing" part. Such algorithms were popularized by Hussain and Rea.[14] Both approaches find matched molecular pairs in an unsupervised fashion and have been implemented in software entitled MCPairs.[15] As shown in Figure 1, the fixed and changing parts that are identified by the two methods can be quite different.

With both methods, molecules are grouped into pairs if the changing and fixed parts satisfy certain criteria. In the MCSS method, the fixed part must account for at least a certain fraction of the total molecule. This is usually assessed according to the number of heavy atoms in the fixed part and in the whole molecule such that $N_{fixed}/N_{molecule}$ (where N is the number of heavy atoms contained in the indicated fragment), must be less than the cutoff styled f_{MCSS}. Thus, for the example in Figure 1, N_{fixed} is 15. Similarly, $N_{molecule}$ is 17 for A and 16 for B and the ratio is 0.88 and 0.94 for the two molecules respectively. In the F+I method, an upper limit on $N_{changing}/N_{fixed}$ is sometimes used to define the molecules to pair.[14] This ratio is not bounded and therefore is not easily analysed. A similar effect can be achieved by requiring the fraction of the molecule represented by the changing part $N_{changing}/N_{molecule}$ to be below a selected limit, called f_{F+I}; this is the approach taken in the work reported below. In both methods, the size of the changing part, $N_{changing}$, can also be restricted to avoid having very large changes in structure (which would otherwise be allowed in larger molecules).[14,16–22] The most commonly applied such limit is that $N_{changing}$ should be less than 10 but 12 and 13 have also been proposed. A few reports suggest

that the change in the number of heavy atoms between the two molecules in a pair should also be limited.[9,23]

In the WizePairZ approach,[13] the chemical environment (the context) adjacent to any structural changes is encoded and this has been adopted in the work described below for matched pairs identified by both the MCSS and F+I methods. The simplest environment is the first atom at each attachment point to the changing part (context level 1), and the largest environment encompasses up to 4 atoms (level 4) from each attachment point (Figure 2).

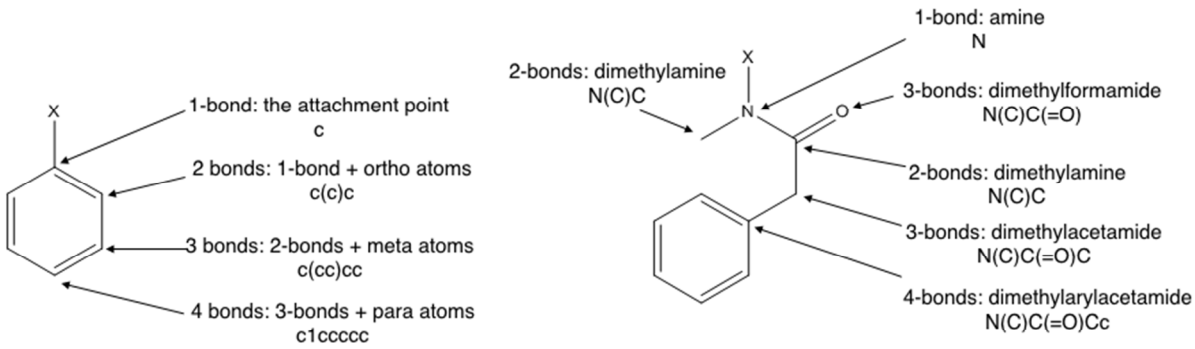

Figure 2. The definition of chemical context for a matched molecular pair.

During the course of the development of the SALT database, it became clear that although the two algorithms have been adopted quite widely, optimal settings have not been reported. Previous work has suggested settings that seem reasonable (such as $N_{changing}/N_{molecule}$ should be less than 0.5 or 0.33)[9,17–21,23] but they have not been verified. These settings have an impact upon the output that is obtained and as such, they are worth optimizing. It became clear that the best way to investigate these settings would be to systematically vary them and probe the effect upon the output. The ChEMBL database of publicly available data has been used to obtain the

optimum settings described below, this permits the data to be presented in full but yields the same conclusions as the analysis of the SALT database.[24]

A positive mean change for a set of matched pairs has previously been found for a small number of pairs that, when expanded by the preparation and testing of further examples, actually has a negative mean change.[3] This suggests that the mean that was obtained might have been positive only because of the sets of pairs that happen to have been made and tested first. To minimize this problem, it is useful to consider that matched molecular pairs are analogous to coin flips where a coin landing heads means that the structural change caused an improvement in the measured property, whereas the coin landing tails means the measured property changed in the undesired direction (Figure 3). When matched pairs are considered as coin flips, the number of actual successes (coin landing heads) can be evaluated for signs of bias. If the coin consistently lands higher or lower than would be expected due to chance, then after a large number of coin flips, there will come a point where the probability of an observed outcome is sufficiently small for an unbiased coin that the observed results are evidence for bias. In the matched pairs analogy, a bias corresponds to a real effect caused by the structural change: a mechanism exists linking the structural change to a property change. This can be viewed as proposing the null hypothesis that a particular structural change has no observable effect, and in order to reject the null hypothesis at a 95% confidence level, the outcome is compared with a series of random coin flips. The 95% CI defines a range of values that has a 95% likelihood of containing the population mean. If the CI contained 0.5, which would be a true mean for a fair coin, the null hypothesis cannot be rejected, or else, if 0.5 was outside the 95% CI, it can be rejected (at that level of confidence) and a real effect has likely been detected.

A great advantage of viewing matched pairs as coin flips is that it readily permits out of range data to be incorporated into the analysis. If one compound is measured as being highly active and out of the range of the assay while its paired compound is less active but in the range of the assay, this is valuable information; similarly if one of the compounds is sufficiently inactive to fall outside the assay. Such examples are highly informative and might include some of those pairs sometimes classed as activity cliffs.[25–27] The smallest number of coin flips for which 0.5 is outside the 95% CI is 6 when all 6 are heads or all 6 are tails (Table 1). This means that in order to make confident assumptions about whether a structural change is indeed changing a property of interest, at least 6 MMPs are required and in all 6 pairs the value of the measured property has to change in the same direction.

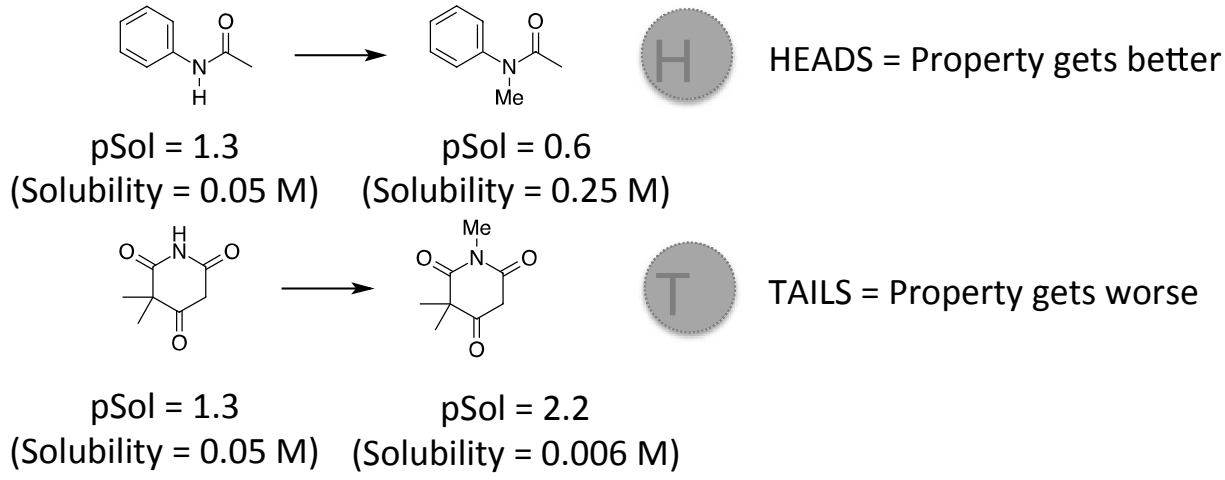

Figure 3. Matched pairs viewed as coin flips.

Table 1. All possible outcomes for flipping a coin 6 times, together with the 95% CI boundaries.

Number of heads	Fraction of heads	Lower 95% confidence bound	Upper 95% confidence bound	Reject the null hypothesis?
0	*0.00*	*0*	*0.459*	*Yes*
1	0.15	0.004	0.641	No
2	0.33	0.043	0.777	No
3	0.50	0.118	0.881	No
4	0.66	0.222	0.956	No
5	0.83	0.358	0.995	No
6	*1.00*	*0.540*	*1*	*Yes*

The coin flip analogy is useful because it reduces the likelihood of incorrectly assigning the direction in which a property is most likely to change for any given change in structure. However, it relates only to the direction of change in a property and does not differentiate transformations causing small effects from those causing larger changes in property.

In the following, we describe our findings concerning how best to identify MMPs with the MCSS and F+I methods and show how the two can be combined effectively. Then we describe how best to analyze the output from MMPA and how to use this to extract increased value with a round of further analysis; a turbocharging effect. Previous work by Willett and co-workers in the area of similarity searching has introduced the idea of using the output of one round of analysis to improve subsequent analyses (in their case clustering).[28,29] They used the term "turbo" to describe this approach and we have applied a similar mechanism to improve MMPA. We show how an analysis using the conservative coin flip approach can provide insights that allow the magnitude of change information to be used with increased confidence and that chemically more specific sets of matched pairs can be more reliable guides than more general sets.

Methods

The SALT database includes MMP data from AstraZeneca, Roche and Genentech. The processing of this data is described elsewhere.[12] For analysis involving ChEMBL data,[24] the molecular weight of compounds was limited to be less than 500 Da. Matched pairs were found using the MCpairs software.[15] This provides an implementation of the MCSS pair finding algorithm described by Warner et al. and of the F+I method described by Hussain and Rea.[13,14] The MCSS identified in MCpairs is limited to contiguous atoms (no disconnected substructures are identified). The structural change linking pairs of molecules is encoded by SMIRKS and these are modified to include differing levels of chemical context. The procedure for the logistic regression analysis is provided in section S4 in supporting information. Statistical analysis was performed in R, confidence interval ranges for binomial probabilities used the exact method and graphs were prepared in ggplot2 and in Vortex.[30–32]

The data that were obtained in the preparation of the SALT database were used to identify improved ways to analyze the matched pairs output. In this case, the pairs were identified using the MCSS approach with settings as described by Warner et al.[13] The number of experimental results included in the MMPA and the resulting number of unique structural transformations that were identified in each set of data is given in Table S1 in the supporting information. "hERG" refers to block of the hERG-encoded potassium ion channel, reported as a pIC_{50}.[33] These data are referred to as hERG throughout. Human hepatocytes are used to measure the metabolic intrinsic clearance rate, reported as the volume of solution (of a fixed concentration of the substrate) that can be cleared per minute by a million cells.[34,35] This value is transformed to give $\log_{10}(Cl_{int})$. Compounds with higher values are metabolized more quickly. The human hepatocyte data are

referred to as "heps" throughout. Human microsomes are also used to study the metabolic intrinsic clearance rate but not in a whole-cell system.[36,37] These measurements are of the volume of solution (of a fixed concentration of the substrate) that can be cleared per minute by a milligram of the microsomes. This value is transformed to give $\log_{10}(Cl_{int})$ and referred to as "mics". Compounds with higher values are metabolized more quickly. The logD is for partitioning between 1-octanol and water buffered to pH 7.4.[38] Equilibrium solubility is the solubility in M measured when starting from a solid sample.[4,39,40] The kinetic solubility is the concentration in aqueous solution after a solution of the compound in DMSO has been added and stirred for a fixed period of time.[41,42] Both solubility measurements are transformed to their $\log_{10}(1/(\text{solubility}))$ values and referred to as "pSol" (equilibrium solubility) and "kin-sol" (kinetic solubility) throughout. Lower values of both correspond to higher solubility. Inhibition of the 2D6 isoform of the cytochrome P450 enzymes is reported as a pIC_{50} and referred to as "2D6".[43] For several properties, measurements made at different companies have been included in the database. Here, the assays have been analyzed separately and the different companies are indicated with a number in parentheses after the property. The numbers are specific to each property: the company designated as (1) for the logD measurements is not the same as that designated (1) for mics.

Results and Discussion

In this section, we first describe optimized settings for MCSS and F+I methods and show how the two methods produce complementary output. Subsequently, we describe how the effect of making a particular structural change can be very variable and show that this variability is

reduced if the set of pairs is more chemically specific. Finally, we use logistic regression to show how a second round of analysis can use information about set size and chemical specificity to extract more useful insights from the dataset.

Optimizing matched molecular pair finding

Both the MCSS and F+I methods have been applied to three datasets from the ChEMBL database representing an enzyme (EGFR), a receptor (the dopamine D1 receptor) and an ion channel (Cav 3.2).[24] These comprised 1006, 903 and 792 individual compounds respectively. Thus, the total number of possible pairings (if each molecule were paired with every other molecule in the set) would be approximately 1×10^6, 8×10^5 and 6×10^5 for each of these sets in order.

The first step was to compare which molecules are paired by each method and to compute the number of individual pairs. The F+I method can link pairs of molecules according to more than one structural transformation whereas the MCSS method finds a single description for each pair. By identifying unique pairs of molecules according to their Chembl identifiers, the two methods can be compared fairly; each pairing is only counted once. The number of pairs produced by the two methods with different values of f_{MCSS} and f_{F+I} is shown in Figure 4. The number of molecular pairs produced for each setting is plotted on a logarithmic scale and the number of pairs alters over several orders of magnitude as the two parameters are changed. In addition to pairs of molecules, it is also of interest to investigate how many structural differences are

identified. These are shown in Figure 4 as red circles and are close to the number of pairs found; most structural changes are represented by only one pair.[14]

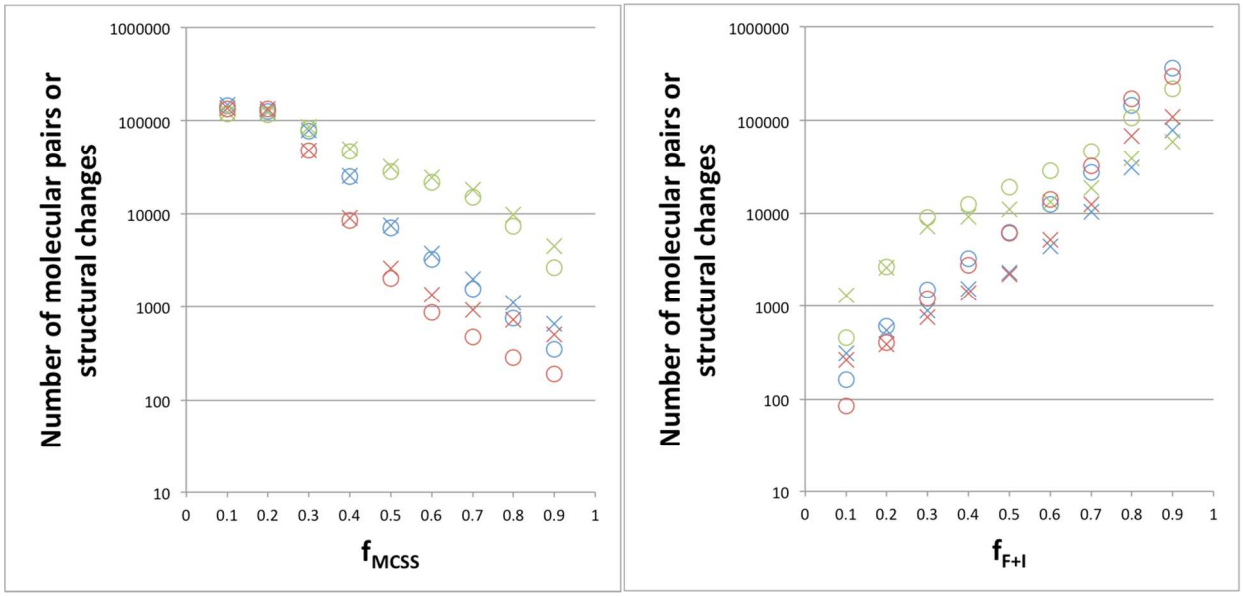

Figure 4. The number of molecules that are found as pairs (crosses) and the number of unique changes in chemical structure (circles) by the MCSS method (left) and fragment and index method (right) for different settings of f_{MCSS} and f_{F+I}. EGF, D1 and Cav 3.2 are colored blue, green and red respectively.

The number of molecules that are unique to each method and the number that are found in common has been computed for each method with the same settings as used to generate Figure 4. From this, the percentage of pairs that are in common between the two methods can be computed. This is indicated by the red segments in the pie charts in Figure 5. The pie charts have sizes that depend on the number of pairs that are found. There is least commonality when either f_{F+I} and f_{MCSS} are both high or both low. The number of pairs found in common is the biggest proportion of the set when high values of f_{F+I} are paired with low values of f_{MCSS} or vice

versa. In general, the ridge of maximum commonality sees $f_{F+I} = 1.1 - f_{MCSS}$. The number of pairs found by both methods represents the greatest proportion of all the pairs found when $f_{F+I} \sim 0.4$ and $f_{MCSS} \sim 0.7$. However, even in the case of the highest degree of commonality, the overlap only represents about half of all pairs found. There remain pairs of molecules that can only be found by one method or another.

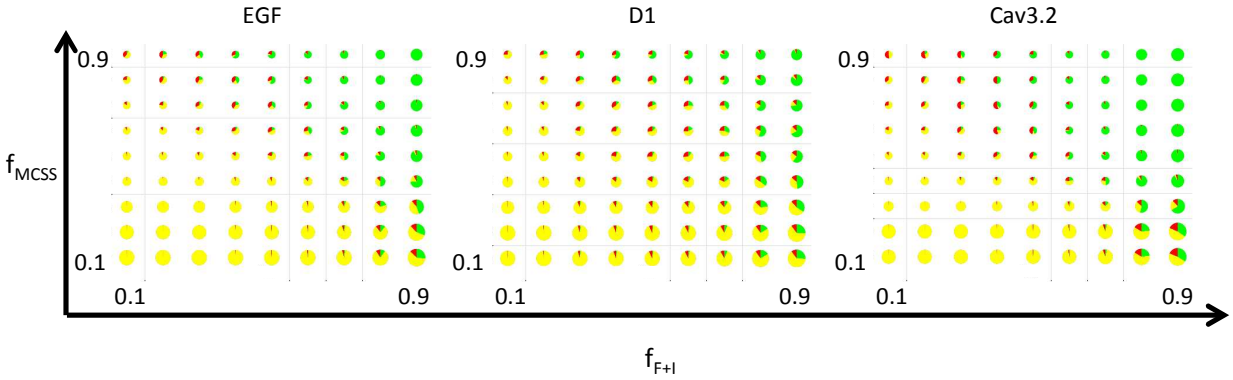

Figure 5. The number of pairs found when both the MCSS and F+I method are applied, with the range of values of f_{MCSS} and f_{F+I} shown and sampled in increments of 0.1. Pie charts are colored as: yellow - pairs found by MCSS only, green – pairs found by F+I only and red – pairs found by both methods.

In order to explore whether this is of chemical significance or not, the set for Cav3.2 in which $f_{F+I} = 0.4$ and $f_{MCSS} = 0.7$ has been examined and one example pair is provided that was found only by MCSS and one pair only found by F+I. The first example (compounds **1** and **2**) shows that when molecules are smaller, they often fail to satisfy the requirements applied for one method or another. In this case (Figure 6), the fixed part and changing part identified by the two methods are different. For the first molecule, the MCSS represents 15 out of 20 heavy atoms ($N_{fixed}/N_{molecule} = 0.75$) and for the second 15 out of 21 heavy atoms ($N_{fixed}/N_{molecule} = 0.71$); in

both cases $N_{fixed}/N_{molecule}$ is above f_{MCSS} (0.7) and so this is an allowed pair. Meanwhile, when the F+I approach is applied, thanks to the differing substitution pattern around the aniline ring, the changing part accounts for 11 of 20 ($N_{changing}/N_{molecule}$ = 0.55) heavy atoms for the first molecule and 12 out of 21 ($N_{changing}/N_{molecule}$ = 0.57) for the second molecule and both values of $N_{changing}/N_{molecule}$ are above f_{F+I} (0.4) and therefore not allowed.

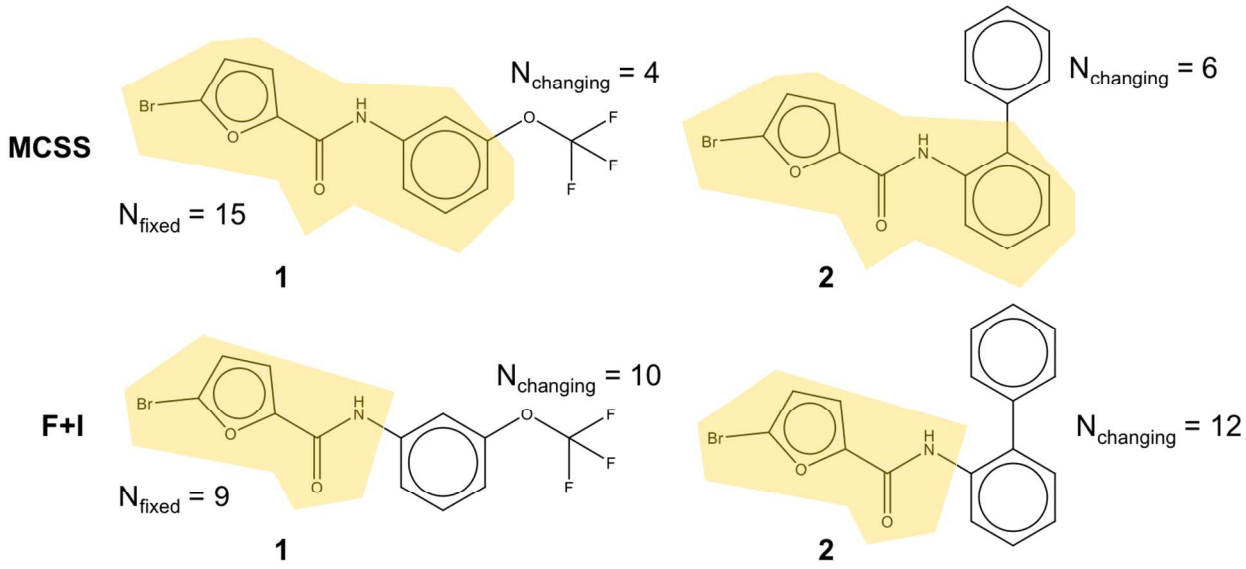

Figure 6. Example of a pair that is only identified by the MCSS method. The fixed part that is found by each method for the pair is highlighted.

In the second example pair (Figure 7), compounds **3** and **4**, found by the F+I method but not the MCSS method, the maximum common substructure (which must be contiguous in the present implementation) is 15 of the 32 heavy atoms ($N_{fixed}/N_{molecule}$ = 0.47) and therefore below the f_{MCSS} cutoff of 0.7. When F+I is applied, the changing part is the central moiety and accounts for 11 out of 32 heavy atoms ($N_{changing}/N_{molecule}$ = 0.34) in both molecules in the pair, which is below the f_{F+I} limit of 0.4 and so they are considered a pair. In this case, the difference in behavior is linked to the details of how pairs are identified and how the two cutoff values interact with the

method; the effect of the cutoff is sensitive to the size of the molecules involved. Neither of these pairs (**1**+**2** or **3**+**4**) is chemically unreasonable and so it must be concluded that using both methods is likely to increase the chances of finding all pairs worth considering.

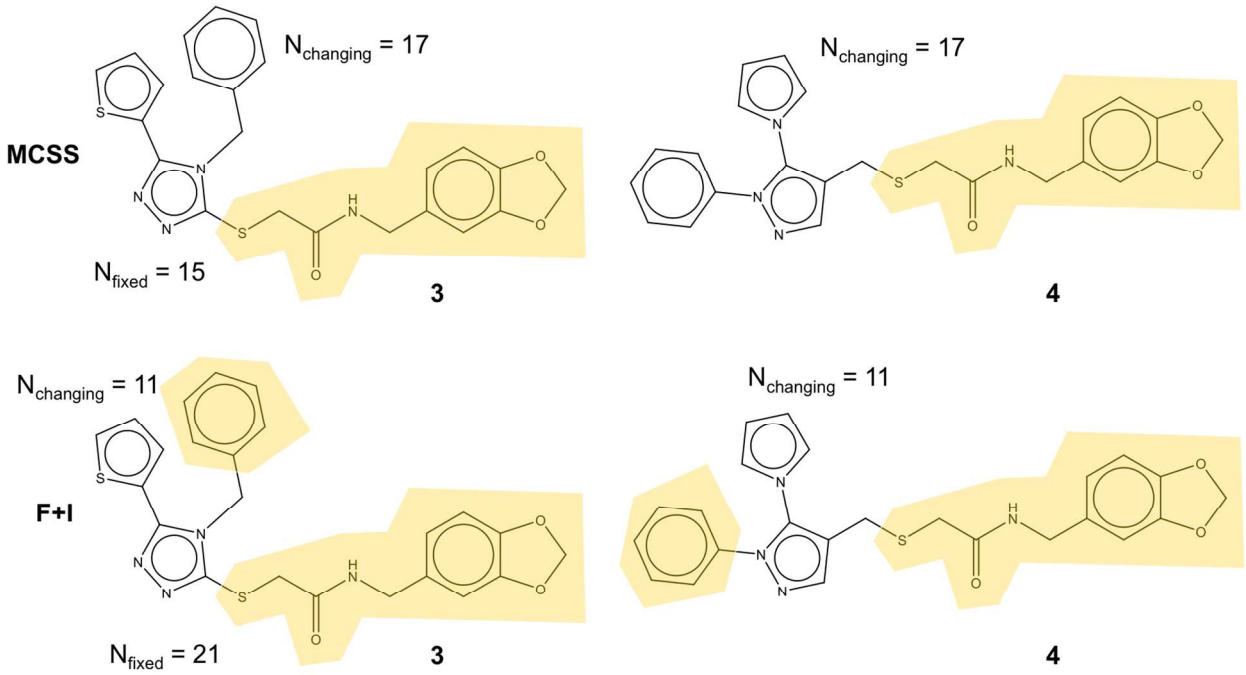

Figure 7. Example of a pair that is only identified by the F+I method. The fixed part that is found by each method for the pair is highlighted.

An extra limit can be imposed based on the number of heavy atoms in the changing part. This limit, $N_{changing}$ is often set to 10 and this value has been applied in the present work.[14,16,17] In order to visualize the effect of this limit, the proportion of pairs that are found by each method is represented as an array of pie charts like those in Figure 5 in section S2 in the supporting information. When the size of the changing part is limited to 10 heavy atoms for the MCSS method, the F+I method contributes most pairs; the overlap is still maximized in the same

regions of the plot as when there are no heavy atom constraints. In an analogous fashion, when a heavy atom limit is applied to the F+I method,[14,16–22] pairs are almost exclusively found by the MCSS method. When both methods are subject to a heavy atom count limit for the changing part, red is a more prominent color suggesting that the two methods behave more similarly. This indicates that many of the transformations that are found only by one of the methods involve large changing parts. As discussed below, MMPA relies on casting the net widely and so imposing a limit of $N_{changing}$ is likely to be detrimental.

There is an alternative way of comparing the methods, according to the number of structural changes that are identified. The MCpairs program encodes structural changes found by the two methods in an identical fashion (as SMIRKS[44,45]) and as such, the transformations found by the two methods can be compared. These are shown in Section S3 in the supporting information. The overlap is smaller than when considering pairs of molecules. This highlights that the two methods will pair the same molecules for different reasons: the F+I method can find changes of large groups (such as substituted phenyl rings) whereas the MCSS method localizes the structural change to the smallest part of the structure that is different between the two molecules in the pair. The two methods are complementary, which suggests that it is advantageous to apply both methods.

The biological data were then added to each pair in order to identify "rules" – the structural changes that exert a real influence. In this case, rules are determined to be statistically significant by treating matched pairs as if they were coin flips. A rule is a set of pairs that support a real effect (the null hypothesis, described above, can be rejected) and the number of rules identified with each of the different settings is depicted in Figure 8. The F+I method produces more rules

than the MCSS method, and the number of rules increases monotonically as f_{F+I} increases. The MCSS method is less sensitive to the value of f_{MCSS} that is selected. The number of rules is maximum when f_{MCSS} is lowest. If each structural change that is identified is a candidate to become a rule then it is clear that for both methods, when more candidates are considered, more rules are found. This suggests that some of the rules might arise simply by chance. To test this possibility, the datasets were reanalyzed but instead of assigning each matched pair as an increase or decrease according to the data for the two compounds, it was instead assigned randomly – treating each pair as an unbiased coin flip. The number of rules that are produced in this instance was computed and then the process repeated three times and the average of these three sets of random outcomes computed. These are shown in hatched bars for the MCSS and F+I method. More rules are found with random data for settings that produce more rules when the real data is used but in all cases, many fewer rules are found using the random approach. In general, the number of rules in the random data is about 20% of the number of rules produced with the real data for both methods. On average, the random set produces a number of rules that is the smallest proportion of the real rules when f_{MCSS} is 0.3 (11.3 %) or when f_{F+I} is 0.9 (8.6 %).

The effect of combining output from the MCSS and F+I methods was also investigated. The number of rules found for each combination is shown in the bottom half of Figure 8. These broadly follow the trends shown for the individual methods but the number of rules found when the two methods are combined has a maximum value that is higher than the sum of the number of rules found by each method alone. The two methods can contribute extra examples (when their output is encoded in a consistent fashion) and by so doing permit statistically significant rules to be found that cannot be found by either method alone. The analysis of random outcomes

shows the same trends as for individual methods. When the three datasets are compared, the proportion of random rules is consistently lowest (below 10 %) when f_{MCSS} is 0.4 and f_{F+I} is 0.9, close to the values for each method alone. These are notably different from the settings that maximize the commonality between the compounds found as pairs by the two methods (see above, where $f_{MCSS} = 0.7$ and $f_{F+I} = 0.4$). This suggests that the methods work best together when they identify complementary but different pairs.

Increasing the number of rules that are considered requires more computational resources. The F+I approach can rapidly require very large storage as the number of compounds increases, particularly if many fragmentations are permitted; a combinatorial explosion can occur. By contrast, the MCSS method can take substantially longer thanks to having to perform many more maximum common subgraph identifications (the number of comparisons scales as $\sim N^2$).

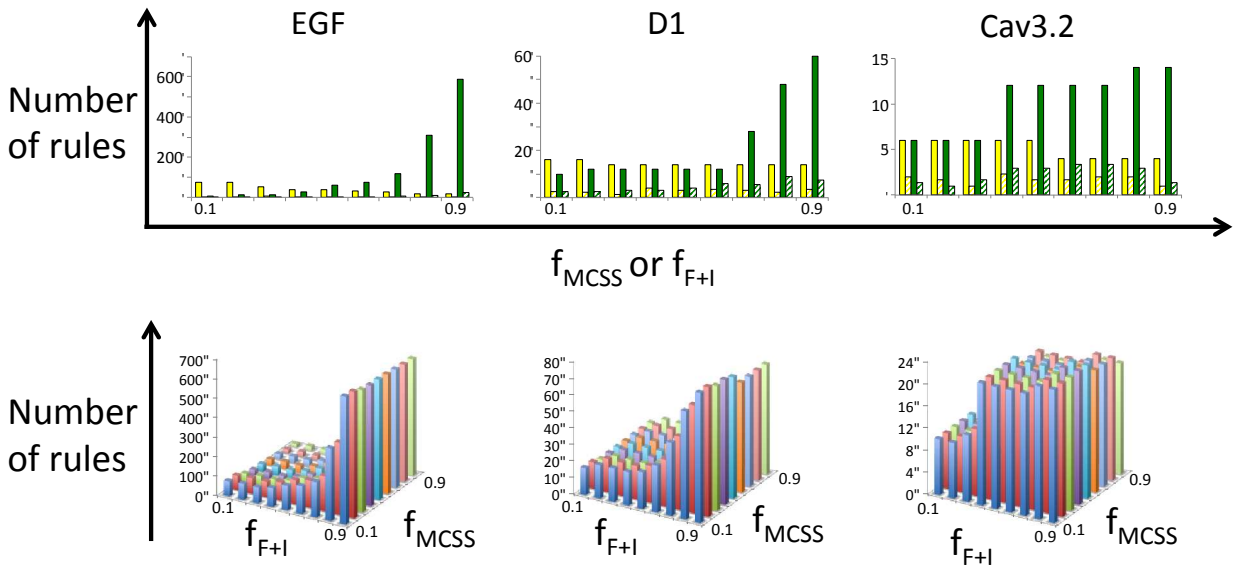

Figure 8. The number of rules found when applying the MCSS (yellow) or F+I (green) methods alone (top) using either real data (solid bars) or random coin flips (hatched bars). The number of

rules found when applying both methods is shown (bottom) for different combinations of f_{MCSS} and f_{F+I}.

MMPA output: variability in the effect caused by a particular structural change

Turning to how best to analyze the output of MMPA, two aspects have been examined: the magnitude of the change in properties and the level of chemical specificity. In this section, sets of pairs were identified using the default MCSS settings described by Warner et al.[13] When a pair is identified, the change in property, referred to as the delta value (Δ) is computed. For example, this might be the pSol difference for each of the pairs shown in Figure 3: -0.7 (top row, heads) and +0.9 (bottom row, tails).

The SD describes how wide the distribution of Δ values for any given structural change is. When the variation of the standard deviation with number of pairs is plotted, as in Figure 9, it becomes clear that the standard deviation is not constant: it increases steadily for small sets of pairs before leveling off to a plateau. After this point the variability has been sampled and adding more pairs (performing more experiments) is likely redundant. For instance, for solubility (pSol), the average standard deviation for a set of pairs levels off at 0.7 log units once approximately 20 pairs are in the set. The values of the plateau and the point at which this is reached have been estimated and are shown in Table 2.[46] Others have shown the value of resampling to obtain a better description of the distribution of Δ values,[47] and the SD has been emphasized as a useful statistic previously.[4] Elsewhere, it has been asserted that the SD should decrease as more pairs are added such that the mean value of Δ divided by SD would be a useful statistic to emphasize influential, consistent sets of pairs but this is in contrast to our findings on a much larger dataset

(Figure 9) which see the SD increase or remain the same as pairs are added.[48,49] The importance of a careful consideration of the contributors to SD, including experimental variation and dataset heterogeneity, has also been highlighted and such considerations are doubtless at play in our observations.[50]

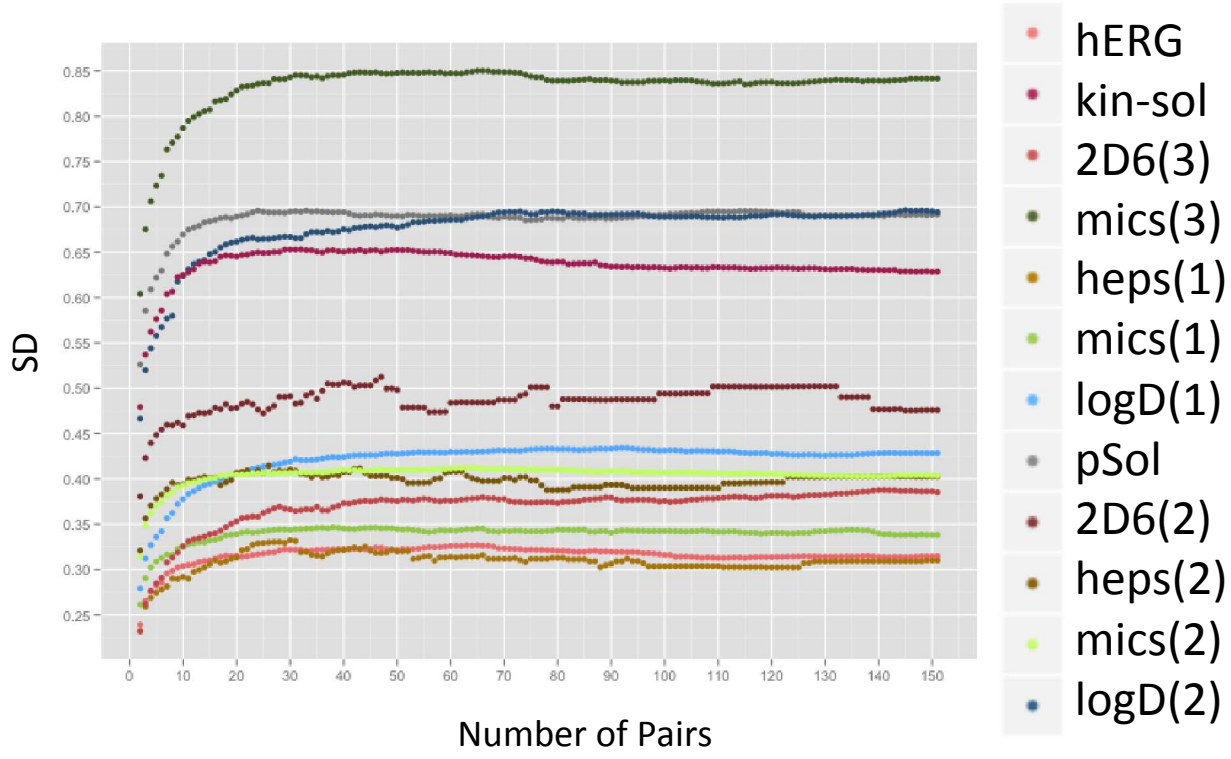

Figure 9. Standard deviation plotted against number of pairs

The SD in a set of matched molecular pairs represents the unavoidable variation in the effect that a structural change exerts. This variation arises for several reasons including the experimental uncertainty in each measurement and the influence of the changing chemical context. This would suggest that more structurally specific groups should have a lower SD. This has been investigated for each of the properties and equivalent plots to those in Figure 9 are shown for different levels of chemical context specification for each property in Figure 10. For all properties, as the chemical context becomes better specified, the plateau for the SD is at

progressively lower values but the number of pairs required to achieve the plateau remains approximately constant. Experimental variation and the residual chemical variation ensure that even when the context is defined at level 4 there remains a significant SD; even chemically well-defined sets have a width that limits how reliable any prediction can be. The properties that see the biggest proportional reduction in SD upon going from context level 1 to context level 4 are logD(1) and mics(1) while mics(3) sees little change.

Table 2. The level at which SD for each property reaches a plateau and the corresponding number of pairs.

Property	Plateau value for SD	Number of pairs to reach plateau
hERG	0.31	25
heps(1)	0.31	15
heps(2)	0.40	10
mics(1)	0.34	24
mics(2)	0.41	32
mics(3)	0.84	29
logD(1)	0.43	62
logD(2)	0.69	66
pSol	0.69	21
kin-sol	0.64	12
2D6(2)	0.49	27
2D6(3)	0.38	65

MMPA output: logistic regression provides settings for the reanalysis of a set of matched pairs

Chemical specificity brings the benefit of reduced variation but the penalty of smaller set sizes. This trade-off was investigated using logistic regression. The analysis was first performed on the hERG dataset and this is used to explain the process before results from other properties are provided. The matched pairs from the hERG dataset were first analyzed using the coin flip

approach. Those sets causing an increase or decrease in inhibition are styled INC or DEC respectively. The remaining sets were assigned to the class No Effect Determined (NED); this includes structural changes that cause no change in hERG inhibition and those that do change the property but for which there are not yet enough data.

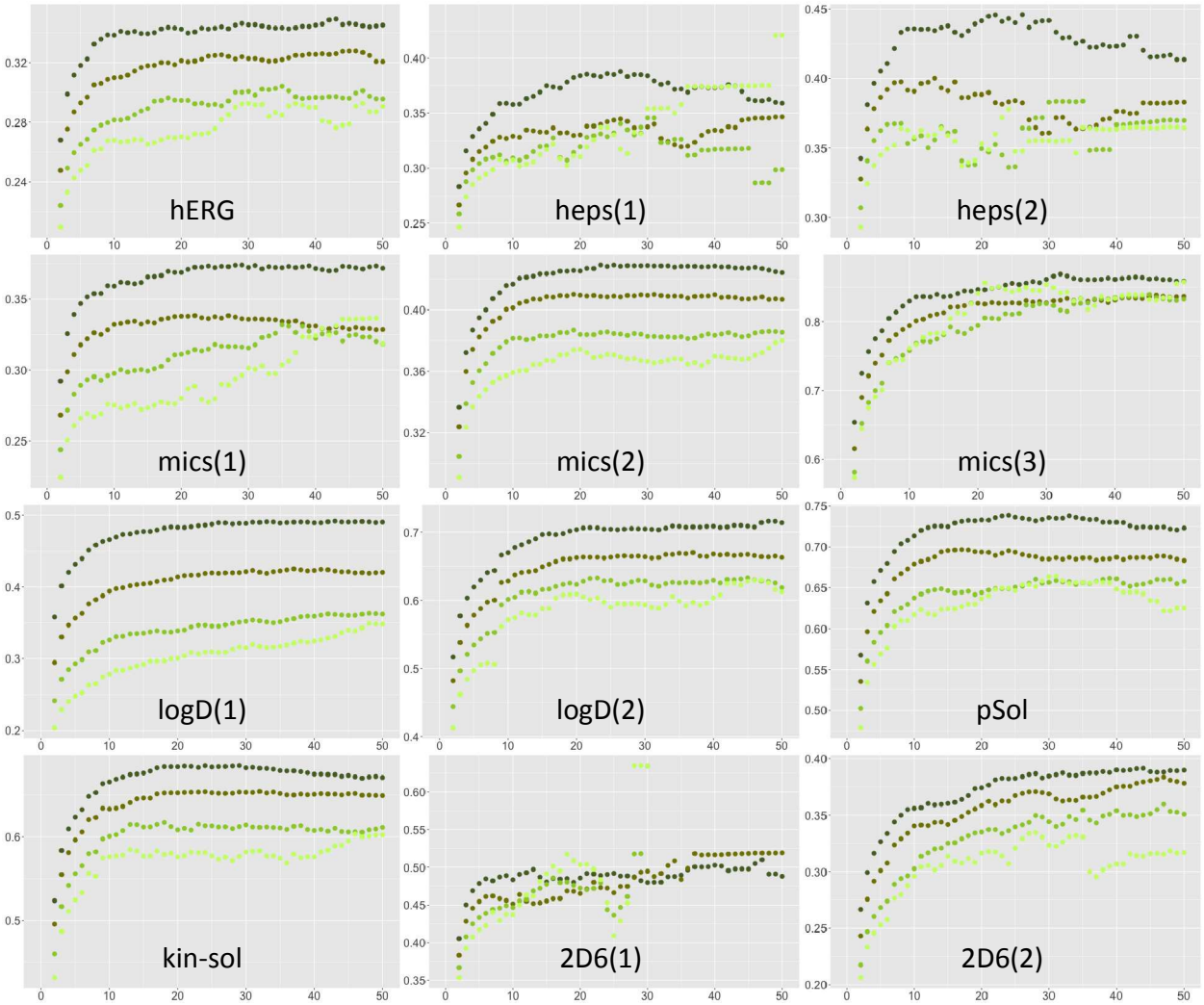

Figure 10. The variation of SD with the number of pairs when subset according to the level of specification of the chemical context. The darkest colored points are for context level 1, the lightest for context level 4 (as defined in Figure 2).

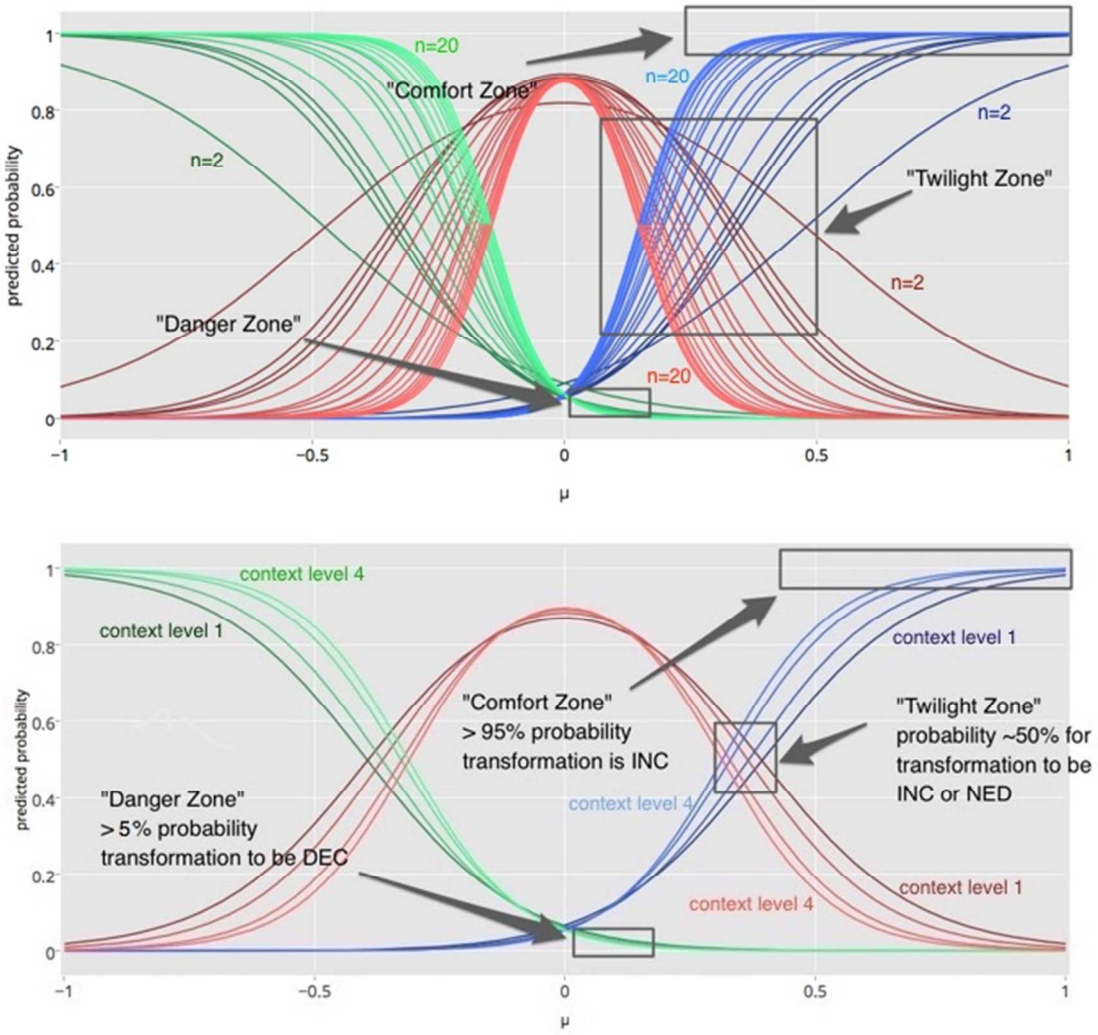

Figure 11. The probability that a set of N pairs with mean value μ will evolve into a set of pairs classified as INC (blue), DEC (green) or NED (red). In the plot at the top groups are subset according to the number of pairs included, the coloring starts dark for N=2 and becomes progressively lighter as N increases to 20. In the plot at the bottom groups all have N=5 and are subset according to the chemical context definition, the coloring starts dark for context level 1 and becomes lighter as it increases to 4.

Having assigned each set to one of three classes (INC, DEC, NED), the measured difference Δ for each pair of molecules in the set was examined. The variation in the average of Δ as pairs are added to a set tracks how sets of matched pairs evolve. If the sequence is Δ_1, Δ_2, $\Delta_3 \ldots \Delta_N$, then the mean of Δ_1 and Δ_2 is μ_2, the mean of Δ_1, Δ_2 and Δ_3 is μ_3. The ordering of pairs in this database is arbitrary (in real datasets it is governed by the order in which compounds are made and tested). Hence, in each set the order of Δs was scrambled and μ_1 to μ_N recomputed. The scrambling is repeated 50 times and the average for each μ computed to give $\overline{\mu_2}$, $\overline{\mu_3} \ldots \overline{\mu_N}$. For each set of pairs, this tracks how the mean value evolves as more pairs are added. These are then averaged over all sets of pairs to give $\overline{\overline{\mu_2}}$, $\overline{\overline{\mu_3}} \ldots \overline{\overline{\mu_N}}$. The evolution along this series corresponds to what happens (on average) for sets of pairs in each class (INC, DEC and NED) as more pairs are added.

Logistic regression analysis links μ with the likelihood of that set ultimately becoming INC, DEC or NED when more pairs are added. These probabilities are plotted against values of $\overline{\overline{\mu_2}}$ to $\overline{\overline{\mu_{20}}}$ in the upper plot of Figure 11. This shows that high values of μ_i are most likely to correspond to INC, low values to DEC and intermediate values to any of the three classes. In the following, we focus upon mean changes that are positive but symmetrical arguments concerning negative mean values would hold. There are three distinct zones highlighted: 1) the 'Comfort Zone', where the probability of INC is more than 95%, 2) the 'Twilight Zone' where both INC and NED are equally probable and 3) the 'Danger Zone' in which there remains a likelihood above 5% that the set of pairs will ultimately be in the DEC class. When the number of pairs is low, the Comfort Zone is only reached for sets of pairs in which the mean change is very high – it is effectively never reached when there are only two pairs in the set. As the number of pairs in

the set increases, the Comfort Zone expands such that sets with lower mean values are increasingly likely to fall into it. However, this converges such that by the time there are 20 pairs, adding more does not provide a marked benefit; a mean change above 0.31 for a set of 20 or more pairs indicates a structural change that is highly likely to increase hERG binding. The twilight zone varies in a similar fashion: a large mean change is required for sets with few pairs before they become most likely to belong to the INC class. For sets with two pairs, the mean change in pIC50 has to be above 0.48 before INC is the most likely class for the set to belong to. There is convergence as the set size increases to 20 pairs, such that mean values of Δ above 0.15 correspond to INC being the most likely class. Finally, the danger zone extends out to 0.12 when there are two pairs in the set and contracts to 0.01 by the time there are 20 pairs.

Turning to the importance of chemistry, the sets can be subdivided according to the different levels of description of their chemical environment. The values of $\overline{\overline{\mu_5}}$ (the average value of delta for sets of five pairs) are shown in the lower plot of Figure 11 where the sets are divided according to the level of context definition (see Supplementary Video S1 for $\overline{\overline{\mu_2}}$ to $\overline{\overline{\mu_{20}}}$). The four context levels are clearly distinct, supporting an advantage for sets with increased chemical specification: sets of pairs with lower Δ values are more able to reach the comfort zone, avoid the danger zone and pass through the twilight zone if they are at context level 4 than at level 1.

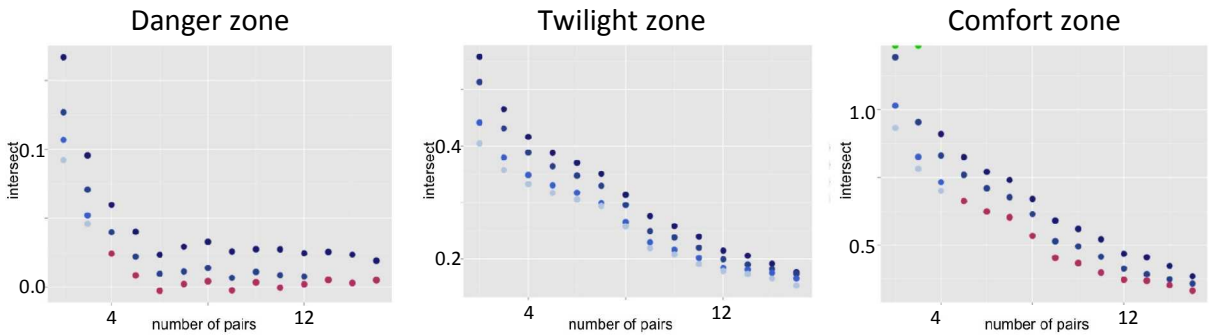

Figure 12. The evolution of the average of the upper edge of the danger zone (left), the center of the twilight zone (center) and the lower edge of the comfort zone (right) as more pairs are added to the set. Points are color coded from dark blue for context level 1 to light blue for context level 4. Each point represents the average of ten samples; when the set of values for two different levels of context definition are not statistically distinct according to a Student's t-test, the points are merged and represented by an average of all the contributing points; merged points are shown in purple.

The analysis of the hERG data is summarized by the evolution of the distributions as more pairs are added to the set and according to the level of definition of the chemical context, shown in Figure 12. In the plot on the left, the edge of the danger zone is shown. If a set of pairs has an average mean delta below the value shown, there is a higher than 5% chance that the sign of the mean change is misleading. This shows that as more pairs are added, the danger zone drops from being 0.17 for context 1 with 2 pairs to less than 0.025 once there are 15 pairs. Also, the better defined the chemical context, the lower the risk of falling into the danger zone; even with only two pairs, the edge of the danger zone is at 0.09 if the context is defined at level 4. The lower edge of the comfort zone is shown in the right hand plot and could not be found for sets of 2 or 3 pairs when context is defined only at level 1. The better defined the chemical context or

the more pairs in the set, the lower the comfort zone extends but as more and more pairs are added, the difference between the different chemical contexts decreases. Levels 3 and 4 are indistinguishable when there are 5 or more pairs in a set. Similar trends are observed for the twilight zone, which is shown in the central plot.

These results are illustrated by an analysis of the ChEMBL database hERG IC_{50} values (full set of pairs is provided as supporting information). Plots similar to those shown in Figure 11 were generated from a set comprising 7347 compounds forming 733 matched pairs of which 259, 241, 161 and 72 were assigned to context levels 1-4 respectively. An illustrative set with context level 4 is provided in Figure 13. Exchanging the highlighted fluorine atom with a methyl group increases the hERG binding in all seven cases. This corresponds to molecules bearing a p-fluorophenyl group. When the set is structurally less well-defined, larger sets of pairs are obtained: 18 at context level 2 and 51 at level 1. These larger but less chemically specific sets don't statistically support a real underlying effect. This suggests that while the general effect of exchanging fluorine and methyl is too small to measure easily, it is larger for p-fluorophenyl groups.

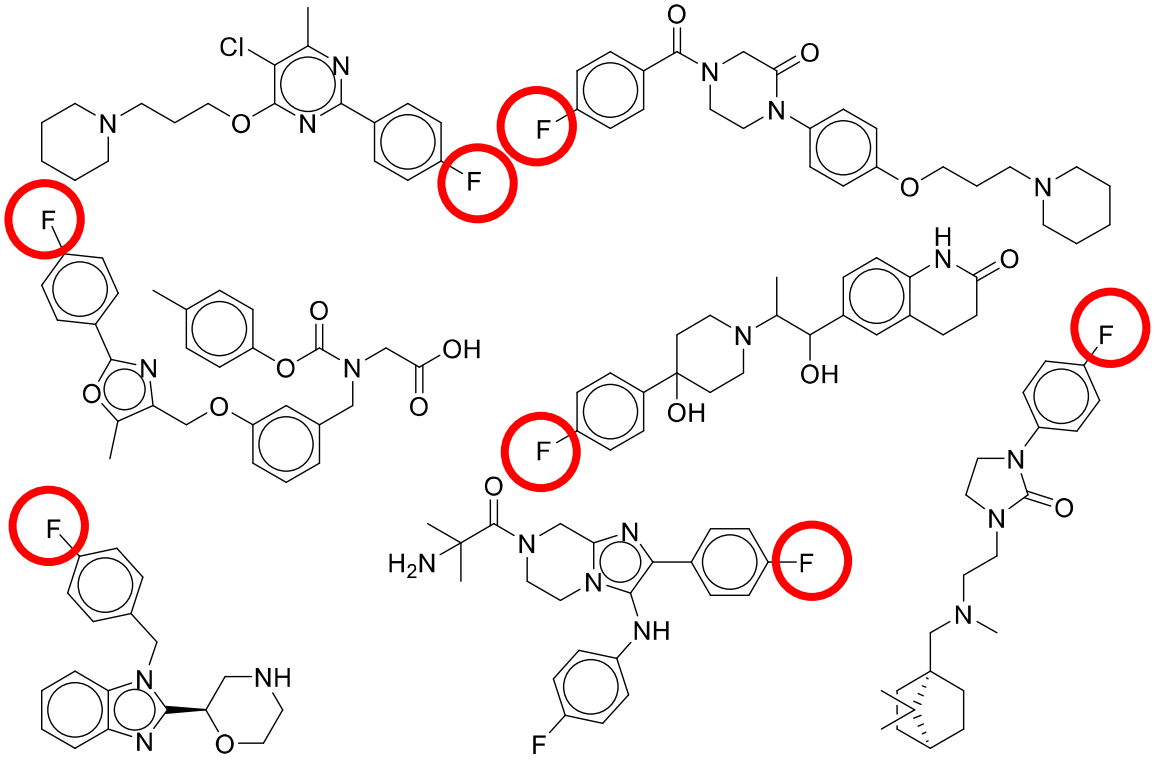

Figure 13. A set of 7 compounds grouped at context level 4 where exchanging the highlighted F for Me increases hERG potency.

It is important to know whether the accuracy of predictions based on sets of matched pairs follow the same trends. This has been tested on the same set of ChEMBL hERG data (Figure 13) by excluding one pair from each set and using all of the remaining pairs in the set to predict the pIC_{50} for the methyl containing compound using the activity of the fluorine containing compound. This was repeated for all pairs and the average error computed. For context level 4 (and 3) the 7 predictions have root-mean squared error (RMSE) of 0.16. The 18 context level 2 pairs give predictions with an RMSE of 0.28 and the 51 context level 1 pairs an RMSE of 0.43. When predicting the outcome for just the 7 context level 4 pairs, the context level 2 set give an RMSE of 0.22 and the context level 1 set an RMSE of 0.30; these are larger errors than for context level 3 or 4 predictions despite being based on more examples. The more chemically

relevant the set of matched pairs used to make predictions is, the smaller the error in the prediction is likely to be. The analysis exemplified for hERG has also been applied to all of the other properties. These are summarized in the supporting information.

Conclusions. The various analyses described above lead to a set of recommendations to improve how MMPA is performed. The two most prevalent methods (MCSS and F+I) for automated identification of MMPs are most effective at finding rules when combined. No matter what settings are employed, neither method alone can find all pairs or rules. The optimum settings see f_{MCSS} set to 0.4 and f_{F+I} set to 0.9. A heavy atom limit on the size of the changing part increases the commonality between the methods but excludes pairs that are worth considering. The first round of analysis for MMPs should use the conservative coin flip approach. This assigns each group of pairs to either increase, decrease or NED (no effect determined). Logistic regression analysis can link the average value of delta with the probability of belonging to each class. This can generate a value for the edge of the comfort zone (95 % chance of belonging to the increase class for positive means or decrease class for negative means) for all combinations of set size (N) and chemical context definition level. The set of pairs can then be reanalyzed. Any set of pairs that exceeds the comfort zone limit corresponding to its N and context level values can be added to the set of rules. Notably, the evaluation of accurate means precludes the use of out of range data but they can be added back in in this reanalysis which only defines lower limits. This turbo charging is particularly important because the number of pairs that represents each structural change follows a Zipfian distribution: small set sizes dominate and any that have less than 6 pairs can not satisfy the coin flip requirements. Analysis of how the standard deviation in delta values varies with set size can identify the innate

level of variability that is to be expected and can identify the number of pairs that are required for good sampling of this variability; making and testing compounds to expand the dataset should avoid expanding sets of MMPs that have already reached this level.

Acknowledgment. JZ is grateful to Medchemica for funding. IL thanks AstraZeneca for funding.

Conflict of interest statement. AGD, EJG and AGL are Directors and Shareholders of Medchemica Ltd.

Supporting Information Available: Numbers of experimental measurements contributing to the SALT database, variation in findings when $N_{changing}$ limit is imposed, variation in number of structural variations found with different methods, example procedure for logistic regression analysis, changes in danger, twilight and comfort zone for SALT database properties. This material is available free of charge via the Internet at http://pubs.acs.orh.

References and Notes.

(1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

(2) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, 2004; 271-285.

(3) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739–7750.

(4) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, 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.* **2006**, *49*, 6672–6682.

(5) Posy, S. L.; Claus, B. L.; Pokross, M. E.; Johnson, S. R. 3D Matched Pairs: Integrating Ligand-and Structure-Based Knowledge for Ligand Design and Receptor Annotation. *J. Chem. Inf. Model.* **2013**, *53*, 1576–1588.

(6) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE: A Matched Molecular Pairs Database for Structure-Based Drug Design and Optimization. *J. Med. Chem.* **2013**, *56*, 5203–5207.

(7) Weber, J.; Achenbach, J.; Moser, D.; Proschak, E. VAMMPIRE-LORD: A Web Server for Straightforward Lead Optimization Using Matched Molecular Pairs. *J. Chem. Inf. Model.* **2015**, *55*, 207–213.
(8) O'Boyle, N. M.; Boström, J.; Sayle, R. A.; Gill, A. Using Matched Molecular Series as a Predictive Tool to Optimize Biological Activity. *J. Med. Chem.* **2014**, *57*, 2704–2713.
(9) Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; MacDonald, S. J. F. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872–1886.
(10) Haubertin, D. Y.; Bruneau, P. A Database of Historically-Observed Chemical Replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(11) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discov.Today* **2013**, *18*, 724–731.
(12) Kramer, C.; Ting, A.; Zheng, H.; Hert, J.; Schindler, T.; Stahl, M.; Robb, G.; Crawford, J. J.; Blaney, J.; Montague, S.; Leach, A. G.; Dossetter, A. G.; Griffen, E. J. *J. Med. Chem.* manuscript accepted. jm-2017-00935p
(13) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* **2010**, *50*, 1350–1357.
(14) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
(15) *MCPairs, version 2.0*; Medchemica Limited: Macclesfield, UK, 2016.
(16) Sushko, Y.; Novotarskyi, S.; Körner, R.; Vogt, J.; Abdelaziz, A.; Tetko, I. V. Prediction-Driven Matched Molecular Pairs to Interpret QSARs and Aid the Molecular Optimization Process. *J. Cheminform.* **2014**, *6*, 1–18.
(17) Kanetaka, H.; Koseki, Y.; Taira, J.; Umei, T.; Komatsu, H.; Sakamoto, H.; Gulten, G.; Sacchettini, J. C.; Kitamura, M.; Aoki, S. Discovery of InhA Inhibitors with Anti-Mycobacterial Activity through a Matched Molecular Pair Approach. *Eur. J. Med. Chem.* **2015**, *94*, 378–385.
(18) Dimova, D.; Bajorath, J. Extraction of SAR Information from Activity Cliff Clusters via Matching Molecular Series. *Eur. J. Med. Chem.* **2014**, *87*, 454–460.
(19) Dimova, D.; Stumpfe, D.; Bajorath, J. Specific Chemical Changes Leading to Consistent Potency Increases in Structurally Diverse Active Compounds. *MedChemComm* **2014**, *5*, 742–749.
(20) de la Vega de León, A.; Bajorath, J. Prediction of Compound Potency Changes in Matched Molecular Pairs Using Support Vector Regression. *J. Chem. Inf. Model.* **2014**, *54*, 2654–2663.
(21) de la Vega de León, A.; Hu, Y.; Bajorath, J. Systematic Identification of Matching Molecular Series and Mapping of Screening Hits. *Mol. Inform.* **2014**, *33*, 257–263.
(22) Wirth, M.; Zoete, V.; Michielin, O.; Sauer, W. H. SwissBioisostere: A Database of Molecular Replacements for Ligand Design. *Nucleic Acids Res.* **2013**, *41*, D1137-43.
(23) Schultes, S.; de Graaf, C.; Berger, H.; Mayer, M.; Steffen, A.; Haaksma, E. E.; de Esch, I. J.; Leurs, R.; Krämer, O. A Medicinal Chemistry Perspective on Melting Point: Matched Molecular Pair Analysis of the Effects of Simple Descriptors on the Melting Point of Drug-like Compounds. *MedChemComm* **2012**, *3*, 584–591.
(24) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
(25) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(26) Hu, Y.; Bajorath, J. Extending the Activity Cliff Concept: Structural Categorization of Activity Cliffs and Systematic Identification of Different Types of Cliffs in the ChEMBL Database. *J. Chem. Inf. Model.* **2012**, *52*, 1806–1811.
(27) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
(28) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. Turbo Similarity Searching: Effect of Fingerprint and Dataset on Virtual-screening Performance. *Stat. Anal. Data Min.* **2009**, *2*, 103–114.
(29) Arif, S. M.; Hert, J.; Holliday, J. D.; Malim, N.; Willett, P. Enhancing the effectiveness of fingerprint-based virtual screening: use of turbo similarity searching and of fragment frequencies of occurrence. In *Pattern Recognition in Bioinformatics: Proceedings. 4th IAPR International Conference on Pattern Recognition, September 7-9, 2009, Sheffield, UK*. , Kadirkamanathan, V., Sanguinetti, G., Girolami, M., Niranjan, M. and Noirel, J., Eds.; Springer: Sheffield, 2009; 404 - 414.
(30) Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer Science & Business Media; Springer-Verlag: New York, 2009.
(31) *Vortex, version 2015.12.46651.25*; Dotmatics: Bishops Stortford, UK, 2015.
(32) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
(33) Sanguinetti, M. C.; Tristani-Firouzi, M. hERG Potassium Channels and Cardiac Arrhythmia. *Nature* **2006**, *440*, 463–469.
(34) Hayes, K. A.; Brennan, B.; Chenery, R.; Houston, J. B. In Vivo Disposition of Caffeine Predicted from Hepatic Microsomal and Hepatocyte Data. *Drug Metab. Dispos.* **1995**, *23*, 349–353.
(35) Ashforth, E. I. L.; Carlile, D. J.; Chenery, R.; Houston, J. B. Prediction of in Vivo Disposition from in Vitro Systems: Clearance of Phenytoin and Tolbutamide Using Rat Hepatic Microsomal and Hepatocyte Data. *J. Pharmacol. Exp. Ther.* **1995**, *274*, 761–766.
(36) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; Macintyre, F.; Rance, D. J.; Wastall, P. The Prediction of Human Pharmacokinetic Parameters from Preclinical and in Vitro Metabolism Data. *J. Pharmacol. Exp. Ther.* **1997**, *283*, 46–58.
(37) Obach, R. S. The Prediction of Human Clearance from Hepatic Microsomal Metabolism Data. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 36–44.
(38) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin. Drug Discovery* **2010**, *5*, 235–248.
(39) Colclough, N.; Ruston, L.; Tam, K. Aqueous Solubility in Drug Discovery Chemistry, DMPK, and Biological Assays. In *Drug Bioavailability*; van de Waterbeemd, H., Testa, B., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2008; pp 7–31.
(40) Yalkowsky, S. H.; He, Y. *Handbook of Aqueous Solubility Data*; CRC Press: Boca Raton FL, 2003.
(41) Zhou, L.; Yang, L.; Tilton, S.; Wang, J. Development of a High Throughput Equilibrium Solubility Assay Using Miniaturized Shake-flask Method in Early Drug Discovery. *J. Pharm. Sci.* **2007**, *96*, 3052–3071.
(42) Lin, B.; Pease, J. H. A High Throughput Solubility Assay for Drug Discovery Using Microscale Shake-Flask and Rapid UHPLC–UV–CLND Quantification. *J. Pharm. Biomed. Anal.* **2016**, *122*, 126–140.
(43) Obach, R. S. Inhibition of Drug-Metabolizing Enzymes and Drug-Drug Interactions in Drug Discovery and Development. In *Drug-Drug Interactions in Pharmaceutical Development*; Li, A. P., Ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2008; pp 75–93.
(44) *SMIRKS*; Daylight Chemical Information Systems Inc.: Santa Fe, NM, USA.
(45) Daylight Chemical Information Systems Inc., SMIRKS - A Reaction Transform Language http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed Feb 16, 2017), 2008.

(46) The plateau was estimated by finding the average value for the SD for sets of pairs with 50 to 150 pairs in them. The number of pairs to reach the plateau is the first time that the SD either equals or exceeds the level of the plateau.
(47) Geppert, T.; Beck, B. Fuzzy Matched Pairs: A Means To Determine the Pharmacophore Impact on Molecular Interaction. *J. Chem. Inf. Model.* **2014**, *54*, 1093–1102.
(48) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
(49) Sheridan, R. P.; Piras, P.; Sherer, E. C.; Roussel, C.; Pirkle, W. H.; Welch, C. J. Mining Chromatographic Enantioseparation Data Using Matched Molecular Pair Analysis. *Molecules* **2016**, *21* , 1297.
(50) Kramer, C.; Fuchs, J. E.; Whitebread, S.; Gedeck, P.; Liedl, K. R. Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. *J. Med. Chem.* **2014**, *57*, 3786–3802.

For Table of Contents use only