Journal of Medicinal Chemistry

PERSPECTIVE

pubs.acs.org/jmc

Matched Molecular Pairs as a Medicinal Chemistry Tool†

Miniperspective

Ed Griffen,[‡] Andrew G. Leach,*[,§] Graeme R. Robb,[§] and Daniel J. Warner[∥]

[‡]Oncology Innovative Medicines Unit, AstraZeneca Pharmaceuticals, Mereside, Alderley Park, Macclesfield, SK10 4TG, U.K.

[§]Cardiovascular and Gastrointestinal Innovative Medicines Unit, AstraZeneca Pharmaceuticals, 30S373 Mereside, Alderley Park, Macclesfield, SK10 4TG, U.K.

[∥]Department of Medicinal Chemistry, AstraZeneca R&D Montreal, Montreal, Quebec, H4S 1Z9, Canada

■ INTRODUCTION

At the very heart of the role of a medicinal chemist or drug designer is the ability to link chemical structure to molecular properties. This is the structure−activity relationship (SAR) or structure−property relationship (SPR). At the start of understanding SAR is the idea that "adding" a particular group "adds" a degree of potency, stability, or other biological property. Many chemists will use this language, for example: "adding a chlorine at the 2-position adds 3-fold potency but lowers solubility by 10-fold". This language is derived from the training of many medicinal chemists being based in synthetic chemistry where thinking literally could be to "add" the chlorine. It is a small conceptual step to then think in terms of a virtual transformation where, although it cannot be done in one step synthetically, the two bioactive compounds can be considered as either end of a transformation. Another conceptual step is to then consider all the occasions a particular change has been tried to be examples of the same transform and to ask questions such as "*How much* potency does adding chlorine at the 2-position generally give?" and "*What is the general effect* on solubility?". The equivalent but systematic approach is to describe the initial compound and the new compound as a "matched pair" linked by the addition of a chlorine at the 2-position and then to look at the statistics of the effects of this structural change. Six years ago one of our colleagues coined the phrase "matched molecular pairs analysis" (MMPA) to describe any systematic method of identifying matched molecular pairs from a set of compounds and determining the property change associated.[1] One of the key advantages of MMPA over other data analysis and modeling techniques is that it deals directly with chemistry and measured data, ensuring clear interpretation of the results. In recent times MMPA has been usefully applied by numerous workers in a range of disciplines within both academia and the pharmaceutical industry.

Drug discovery projects frequently discover that a particular structural change causes a change (for good or ill) to a property of interest, as described above. Two contrasting fates often follow. In the first case, the effect is described to others and what had been an observation in one series is passed on as if a general rule of thumb, whether general or not. In the second case, the effect is not disclosed beyond the original project and the knowledge, either explicit or tacet,[2] does not get passed on. MMPA provides a means to both test the generality of rules of thumb and to extract understanding that might otherwise be lost.

The four authors of this Miniperspective have been involved in developing and implementing some variations on the original theme of MMPA and felt this to be the time to highlight some of these and to point toward future developments. All of us are involved in the process of drug design, and it is within this context that we will set our comments. In outline, we shall introduce matched molecular pair analysis and highlight why it fits so well with the kind of problem solving that is required in the process of drug design. This will be followed by some comments concerning when the application of the methodology is appropriate. Then in a section intended for those with more computational interests, we will describe the various ways that have been devised for identifying such matched pairs. Finally, we will present our expectations for where this approach might be developed in the future.

Throughout the paper, we have elected to use a publicly available data set to illustrate some of our comments: the ChEMBL database of inhibitors of the epidermal growth factor receptor tyrosine kinase (EGFR).[3] This consists of 2899 measurements of an IC_{50} annotated as a binding measurement on 2348 unique compounds. These measurements are drawn from a number of sources, having been abstracted from the scientific literature. We use examples based on this mixed data set purely for illustrative purposes.

■ DEFINING MATCHED MOLECULAR PAIR ANALYSIS AND ITS RELATIONSHIP TO OTHER ESTABLISHED APPROACHES

Matched pairs have generally been defined similarly to the definition proffered 5 years ago by one of us as "molecules that differ only by a particular, well-defined, structural transformation".[4] It is worth noting that while this label emphasizes the two molecules that are related, others have emphasized the transformation linking them.[5] Although the labels vary, the key concept is that the effect of small structural differences between molecules is more easily predicted than the absolute value for the activity or property of each molecule, which is the approach taken in the field of quantitative structure−activity relationships (QSAR) and quantitative structure−property relationships (QSPR). Viewed like this, MMPA and QSAR/QSPR are related in a similar fashion as free energy perturbation methods are to the scoring of bound poses in the field of prediction of

Received: April 14, 2011
Published: September 22, 2011

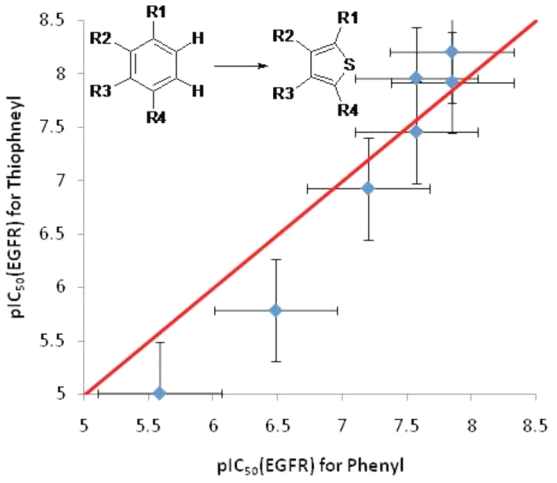

Figure 1. Matched molecular pairs in the EGFR data set that differ by transformation of phenyl to thiophenyl. On the *x*-axis is plotted the pIC_{50} value of the phenyl compound and on the *y*-axis the value for the thiophenyl. The red line is the 1:1 line. Error bars correspond to 3-fold variation in the IC_{50} and indicate typical experimental uncertainty.

protein−ligand binding energies. In that arena as here, the methods focusing on differences as opposed to absolute values tend to be more successful. The origin of this difference in performance owes much to the cancellation of errors and to the averaging out of localized effects present in individual cases.

The quest to quantify the effect of structural changes upon properties of pharmaceutical relevance has emerged as a related discipline to the identification of bioisosteres.[6] These can be defined as "the replacement of a part of a bioactive molecule with a substructure that is similar in size and exhibits similar properties."[7] While the emphasis in bioisosterism has usually been to find groups that are unlikely to change the primary binding potency of a compound, the concept applies quite readily to finding molecular transformations that are likely to leave any given property unchanged. Such transformations are naturally and inevitably identified during the course of most matched molecular pair analyses. For instance, when studying matched molecular pairs that occur in druglike compounds, Sheridan found that phenyl is often exchanged with thiophene. In the context of the EGFR data set, all the pairs in which this transformation has been imposed have been identified and are shown in Figure 1. These show that within this data set this particular isostere is valid. For the overall data set, the mean change is −0.12 with a standard error in that mean of 0.16, indicating that an underlying mean change of 0 is not contraindicated. For individual cases, the error bars on the plot indicate 3-fold variation in the IC_{50}, a typical experimental uncertainty; most cases do not show differences beyond what can be accurately determined.

MMPA might also be considered a relation of Free-Wilson analysis.[8] In this kind of analysis a least-squares fit is used to link different R-groups around a molecular core to contributions to a property of interest. The contrast between the two is illustrated in Figure 2. A notional series of compounds is illustrated as a simple substituted phenyl ring. Free−Wilson analysis is able to make predictions for the contribution of each of the five substituents at defined positions: A_1, A_2, A_3, B_1, and B_2. Meanwhile, matched molecular pair analysis is able to estimate the effect of transforming A_1 to A_2 and of B_1 to B_2 and A_2 to A_3 but

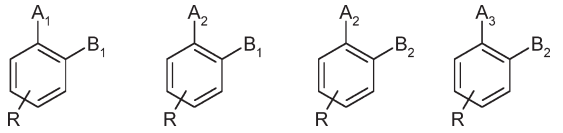

Figure 2. Example illustrating the contrast between Free−Wilson analysis and matched molecular pairs. A series of compounds with defined substitution positions is illustrated as a simple substituted phenyl ring.

not A_1 to A_3; no pairs exist in which two molecules differ only by the change A_1 to A_3. The effect of the last change can only be inferred from the sum of the change caused by changing A_1 to A_2 and that of changing A_2 to A_3. It has been demonstrated that in some cases this sum of pairwise effects is equivalent to the direct transformation.[4] Free−Wilson analysis is a powerful way of understanding the contribution various groups make within a well-defined part of chemical space; the ability to directly link A_1 to A_3 can be viewed as a strength of the analysis. However, if there is substantial interaction between A and B, then the link that is made may be misleading. MMPA would not make that link directly; unlike Free−Wilson analysis additivity of SAR is neither assumed nor required.[9] While there are relationships with the modeling techniques of Free and Wilson, no model is ever built from the data and the link back to measurement remains clear and unambiguous.[8]

In close analogy to the approach of Topliss,[10] MMPA can also be used to derive small sets of substituents to probe a given property. Dossetter studied the groups to which phenyl substituents had been transformed (limited to substituted phenyls and unsubstituted pyridines) in a data set of microsomal metabolic clearance.[11] The range of changes that had historically been brought about by these transformations can set expectations for what might be expected of them in the future, and a representative subset can be picked. This subset should span microsomal clearance in a similar way to the Topliss set of substitution patterns that were selected to span electronic and lipophilic properties. This can guide the design of libraries to probe the changes in clearance that might be expected by modifying certain parts of a chemical series.

QSAR or QSPR methods are also frequently used within the pharmaceutical industry and allow properties of the whole molecule (measured or calculated) to be related to the activity or property of interest.[12,13] Though many successful QSAR and QSPR models have been reported, these models have inherent disadvantages. By their very nature they tend to find the most generalized relationships that smooth over interesting substructural effects[14] and tend to be limited in the precision of prediction. Such generalizations in the prediction of primary potency, and other properties, are often insufficient for guiding design in a lead optimization phase. Precision and accuracy are further limited by the quality of descriptors used in the model. Although many thousands of descriptors can be generated, the value of each is not always clear and can lead to misleading relationships and chance correlations, depending on the data and modeling technique used.[15] Descriptors frequently describe chemical structure incompletely, and so information about the structure−activity relationship is lost.

To illustrate the differences, the EGFR data set has been treated with a very simple QSAR approach and with MMPA. One of the most commonly occurring transformations in the EGFR data set is that of converting a *m*-bromoaniline group into a

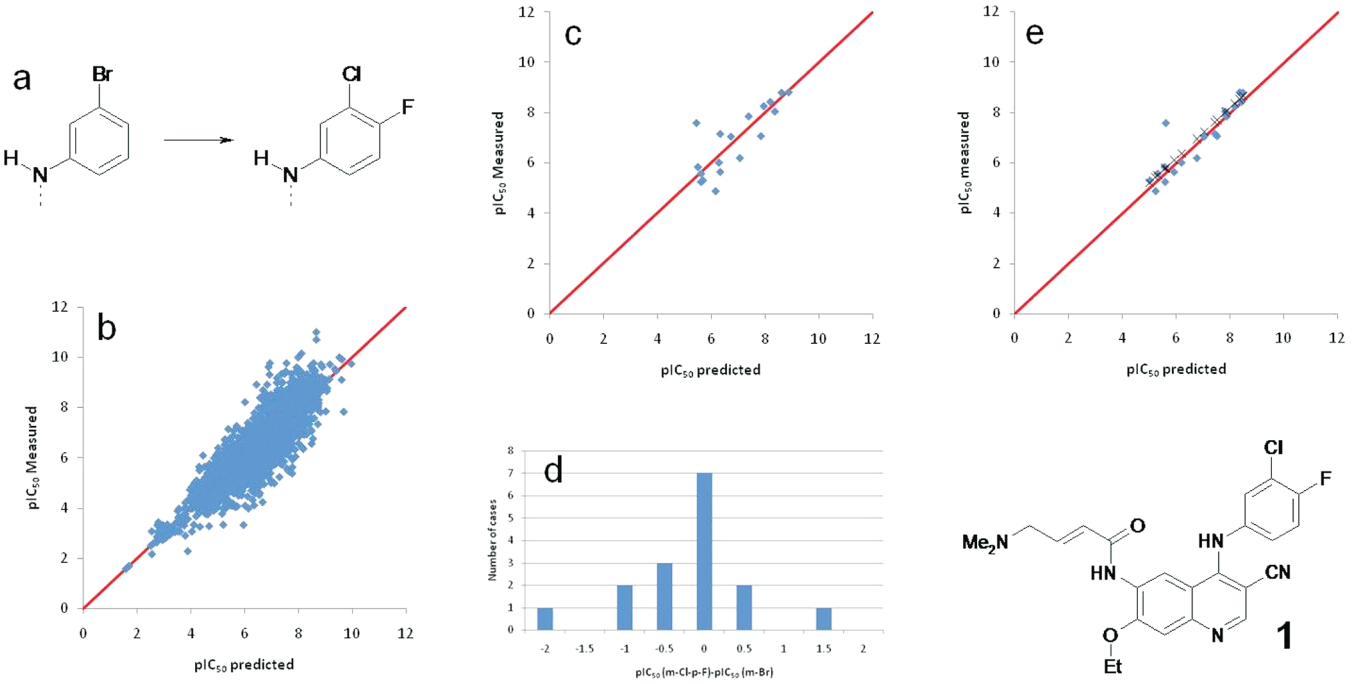

Figure 3. (a) Structural change being studied for its effect upon pIC_{50} against EGFR. (b) Measured values for a training set of compounds plotted against predictions from a simple least-squares QSAR model based upon 193 descriptors for 2287 compounds. (c) Measured values plotted against predictions for a test set of 18 compounds all containing the *m*-chloro-*p*-fluoroaniline using the QSAR model built to generate (b). (d) Distribution of changes in measured values between 16 training set molecular pairs differing only by the transformation of *m*-bromoaniline to *m*-chloro-*p*-fluoroaniline. (e) In blue are the measured values of the same test set as in (c) plotted against the prediction based on matched pairs, assuming that all compounds will experience a mean change of −0.18. In black crosses are plotted the measured values for the *m*-bromoanilines upon which these predictions are made.

m-chloro-*p*-fluoroaniline, as shown in Figure 3a. There are 34 occurences of this matched pair in the data set. Pairs were assigned a random number between 0 and 1, and where the value was less than 0.5 (18 pairs in this case), the *m*-chloro-*p*-fluoroaniline analogue was assigned to a test set. A training set was built in which these 18 *m*-chloro-*p*-fluoroaniline containing compounds were excluded. A QSAR model was built for the 2287 remaining compounds using 193 standard descriptors (40 compounds were not amenable to calculation of all of the descriptors, and 3 did not have IC_{50} data) as computed by the in-house c-lab tool.[16] The model was built using a simple least-squares fitting approach in the JMP package.[17] An excellent fit is obtained for both the training set and the test set, as shown in parts b and c of Figure 3. The training set achieves an R^2 of 0.78 and RMSE of 0.66, while the 18-member test set achieves an RMSE of 0.74. All of these are perfectly reasonable values, suggesting a well-behaved QSAR model. By contrast, the 16 matched molecular pairs in the training set show the changes in pIC_{50} illustrated in Figure 3d. The distribution has a mean of −0.18 and a standard error in that mean of 0.18 and standard deviation of 0.73. Adding −0.18 to the pIC_{50} for each of the 18 *m*-bromoaniline containing compounds corresponding to the test set led to the predicted pIC_{50} values plotted in blue in Figure 3e. The points marked with black crosses are the corresponding values for the *m*-bromoaniline from which each prediction is projected. The predictions are generally excellent; the RMSE for the test set is 0.55. The performance of the matched pairs prediction in the test set is unexpectedly better than in the training set, where the RMSE is 0.70. Figure 3 e illustrates why matched pairs tends to be more successful in terms of predictivity, as it shows that the challenge has been reduced. The potency of the *m*-bromoaniline compound provides a good starting point for predicting the potency of the transformed compound. Just as important, the outliers away from the matched pairs plot provide insight, as highlighted in many literature studies;[4,18] the outlier in Figure 3e is compound **1**. Knowing that this is an outlier can be a spur to drive hypothesis generation, which in turn leads to design of new compounds or further testing of existing compounds.

Although the overall quality of a QSAR model can be determined, the applicability of the model to specific chemistries cannot easily be assessed. Even if a "good" QSAR model that could accurately predict a given activity was discovered, on its own it can provide no information about what to make in the next cycle of drug design. This "inverse QSAR" problem is one to which MMPA is particularly suited. Changes in chemical structure are linked directly to changes in property; knowing what the desired change in property is, a suitable structural change can be selected. The lessons from the historical data allow medicinal chemists to make the best predictions possible about what particular structural changes will achieve in the future.

■ CONSIDERATIONS WHEN APPLYING MATCHED MOLECULAR PAIR ANALYSIS

There are a number of issues that ought to be considered carefully when applying MMPA. These are grouped together here into those concerning the chemical structures and those concerning the preparation and analysis of the associated data. The principal concern in the area of chemical structure is how the "tradeoff (sic) between specificity and generalizability", as highlighted by Papadatos et al. is achieved.[18] Stated in an alternative

dx.doi.org/10.1021/jm200452d |*J. Med. Chem.* 2011, 54, 7739–7750

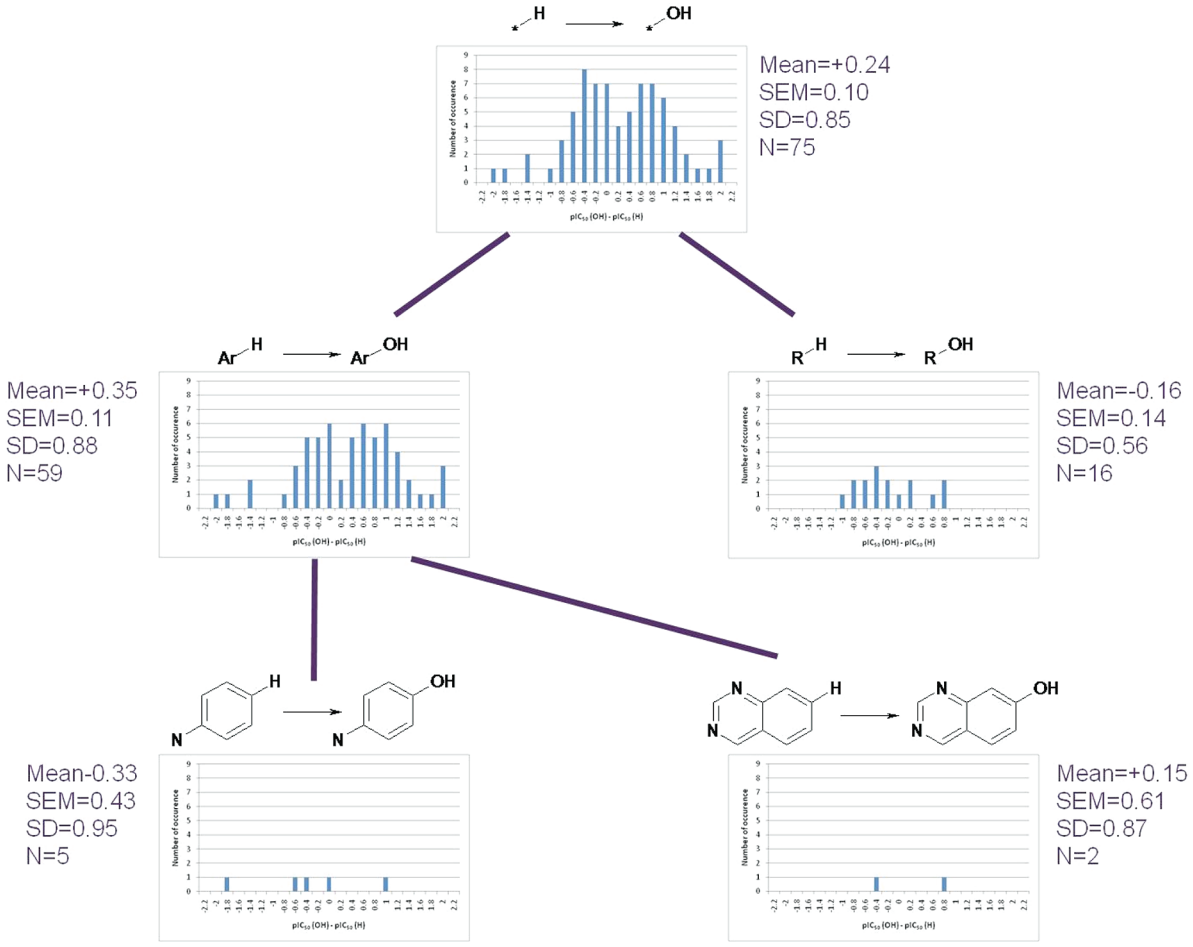

Figure 4. Distribution of changes in EGFR pIC_{50} caused by adding a hydroxyl group in varying contexts with the most general case at the top and some specific positions at the bottom. The distributions are summarized by mean values, standard errors in those means (SEM), the standard deviation (SD), and the number of representative pairs identified (*N*). Ar represents any aromatic carbon linked group and R any aliphatic carbon linked group.

fashion: the more tightly the context of a given structural change is controlled, the smaller the data set of relevant transformations will be and the less structural diversity will be sampled. These two factors conflate to restrict how generally applicable the derived estimates of the change in the property value will be. This is discussed in many of the papers referred to in this article. Here we illustrate the issue with the EGFR data set. In Figure 4 a hierarchy of sets of matched molecular pairs is given. All of the pairs correspond to the addition of a hydroxyl group. At the very top is the set of all such matched pairs. Beneath this are two subsets, one in which hydroxyl is added to an aromatic carbon and the other in which it is added to an aliphatic carbon. The distribution in changes in potency for each of these subsets is given along with the corresponding statistics. These two subsets are statistically distinct from one another and suggest that if addition of hydroxyl is being considered, addition to an aromatic carbon should be preferred over addition to an aliphatic carbon. In the bottom layer, the distributions of changes in potency are given for two of the many possible subsets of the aromatic carbons, one in which the hydroxyl is added in the para position of an aniline and a second in which it is added to the 7-position of a quinazoline. These two sets are more likely to represent the kind of structural change of interest in a lead optimization campaign, but neither has enough representatives to be very general and neither is statistically distinct either from each other or from the more general set of changes caused by adding hydroxyl to an aromatic carbon. While the general rule that addition of hydroxyls is likely to be beneficial for EGFR potency is obtained (top layer), this may not be of use to projects limited for some reason to adding hydroxyls either to aliphatic carbons (middle layer) or to particular aromatic carbons (bottom layer).

When the context is completely proscribed to a particular attachment point on a particular series, the analysis almost reduces to Free−Wilson analysis (see Figure 2 for differences), which is not necessarily applicable or relevant beyond that context. If the context is completely uncontrolled, then the effect of a particular transformation is likely to reflect only simple trends, most notably lipophilicity determined effects.[4,18,19] However the subsetting is performed, a threshold number of molecular pairs must be included in a set before the mean change becomes stable to the inclusion of additional pairs and a significant degree of molecular diversity is required among the molecules contributing to a set of pairs before chemists using the analysis can have confidence that any prediction is likely to be applicable in a general fashion. The first of these concerns is dependent upon the data type being analyzed, whereas the second requires an estimate of the diversity encompassed within

dx.doi.org/10.1021/jm200452d |*J. Med. Chem.* 2011, 54, 7739–7750

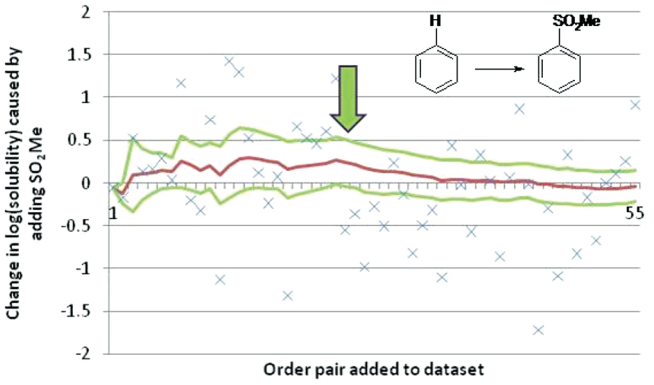

Figure 5. Variation with time in the mean change in solubility corresponding to addition of methyl sulfone to phenyl rings. The individual matched pairs are plotted as blue crosses in the order in which they were added to the data set, and the mean following the addition of each point is plotted in red. The green lines are at 2 standard errors in the mean above and below the mean at each point. The green arrow indicates the point in time at which the data in the publication by Leach et al. were extracted.[4]

a set. This was achieved by manual inspection of the contributing molecular pairs in the publication of Leach et al. but is automated in internal AstraZeneca Web pages by computing the mean Tanimoto difference among the first members of all pairs in a set.[4] These Tanimoto differences can be benchmarked to identify a cutoff that is felt to represent adequate diversity for general applicability. In their study, Gleeson et al. control for diversity by ensuring that each set of compounds contains representatives of at least five clusters, grouped by Tanimoto comparison of Daylight fingerprints, and ensuring that each set has at least 20 pairs.[20] These two considerations of data set size and diversity are too rarely considered in publications in this area. A cautionary note is provided by data related to that originally presented by Leach et al. and shown in Figure 5.[4] The addition of methyl sulfone groups to phenyl rings was reported in that paper to increase solubility. The variation of the mean change in solubility as measurements have been added to the database is shown in the period leading up to that publication and since. It is noteworthy that the transformation now is found to correspond to a small reduction in solubility. Meanwhile the structural diversity, as assessed by mean Tanimoto distance among the phenyl sulfones, has remained fairly constant at ~0.6, having reached that point after about 10 pairs were included. At the time of the publication the mean change in log(solubility) was +0.26 with a standard error in that mean of 0.12; currently the corresponding mean value is −0.03 with a standard error of 0.09. The two mean values are consistent with an underlying population mean in the range +0.02 to +0.15 (see green lines in Figure 5), but the variation in the mean from positive to negative is an extreme example of how temporal variation might dent confidence in the value of the analysis. Such variations are less likely the larger the data set is and the greater the structural diversity represented, but this relationship will depend upon both the structural change being studied and the nature of the data being mined.

A final consideration related to chemical structure that is often neglected is stereochemistry. This can be difficult to treat correctly, particularly if matched pairs are defined by changes at one end or other of a bond to a stereocenter. This is particularly the case in many of the fragmentation approaches discussed below along with MCSS approaches that often use molecular graphs that neglect stereochemical information. For some properties, this is likely to have only a minor influence (lipophilicity for instance), but for others it will have a profound effect upon whether the pairs are appropriately paired or not and hence whether the resulting property change is computed correctly. For all of the reasons outlined in this section, it is beneficial to provide a link to the original data upon which any matched pairs analysis is based such that users can decide for themselves whether the result is relevant. Even rules that appear to be general should be subject to such scrutiny.

The requirements of the data to be analyzed should also be considered carefully. Generally, the output from a matched pairs analysis is either a mean change in a particular property or else the proportion of times that a property has changed in a particular way (i.e., increased vs decreased or active vs inactive). The first of these requires that the property being compared in a pairwise fashion is within the dynamic range of the assay and on a linear scale (and preferably normally distributed). For some properties, such as IC_{50} values, this might involve log transformation to yield pIC_{50} values or equivalents as suggested by Fujita and Ban in their modified version of Free−Wilson analysis.[8,20,21] Properties that are like rates or equilibrium constants should only properly be compared as ratios such that logarithmic transformation should be performed if linear comparisons are to be made, as is the case in MMPA. Logarithmic conversion is also applicable to the analysis of physicochemical and pharmacokinetic parameters, as has been demonstrated by Leach et al. in their study of aqueous solubility, plasma protein binding, and in vivo plasma exposure following oral dosing.[4] A recent paper describes multiplicative transformations, but these result from inappropriately comparing untransformed properties.[2]

Careful consideration should also be given at the outset of what types of data should be paired together for comparison; differences between compounds measured in different assay formats may reveal only differences between assays and may not be related to the structural change. For example, Leach et al. only analyzed solubility data from an assay using solids provided by chemists. Data obtained in assays starting from DMSO solubilized samples were not included. If the need for more data to create larger matched pair sets and increase chemical diversity had been felt, these data could have been included with the risk that the extra variation would add more noise from which a signal might be more difficult to detect. Hajduck et al. compared data from a large set of different assays and end points in a way that is not recommended.[19] Careful consideration should also be given to which sets of pairs should be aggregated together to arrive at summary statistics. Even when data are obtained in comparable assays within each pair, in some cases groups of pairs measured in different assay formats might be sensibly aggregated and in others not. The decision about what to group together should ideally be taken in collaboration with those involved in the design and performance of the experiments. It is with these caveats in mind that our findings based on the EGFR data set are illustrative examples only.

The mean change ought to be presented alongside the standard error in the mean to determine if it is actually distinct from zero (mean ± 2(SEM) should not include 0 to provide 95% confidence that the structural change generally causes a change in the property). This was illustrated graphically by Dossetter in his analysis of microsomal Cl_{int} values where the mean was plotted against the standard error for a set of pairs in which phenyl groups

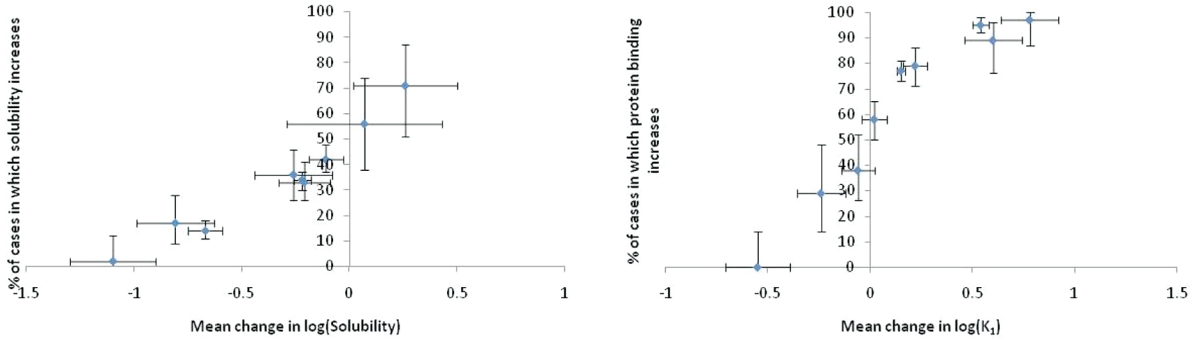

Figure 6. Proportion of cases in which a property increases plotted against the mean change in the property for log(solubility) on the left and log(K_1) for plasma protein binding on the right. Error bars correspond to 2 standard errors in the mean on the *x*-axis, and the binomial probability 95% confidence interval is on the *y*-axis. All data are taken from Leach et al.[4]

are exchanged for alternative substituted aromatic rings.[11] Pair sets that lie above the $y = 2x$ or below the $y = -2x$ line are the transformations satisfying this criterion. The standard deviation should also be considered in order to provide the likely range of values in the change in property to be expected when a transformation is performed. Two standard deviations about the mean ought to encompass the actual value of the change in 95% of cases. Although testing that a mean is distinct from zero might provide confidence that the change is real, when this statistical test is not passed, the values still facilitate the presentation of the relevant information. In such circumstances, it could be concluded that a bioisostere has been identified or else that more measurements are required to enhance the reliability of the mean value.

Many assays have limited dynamic ranges; values outside of the dynamic range cannot be compared to other values (within or without the dynamic range). In this circumstance, it may be that many pairs of interest might be excluded when computing the mean change; oftentimes these are the most interesting pairs because they involve the most substantial changes in a property. This kind of transformation is likely to include the discontinuous changes identified by Wasserman and Bajorath and the switch transformations highlighted by Keefer et al.[2,22] In these cases, it may be appropriate and useful to instead analyze the number of times that a property changes in a particular direction rather than the degree of change. These proportions ought to track with the mean difference. This was shown by Leach et al. for solubility and plasma protein binding and illustrated in Figure 6.[4] If the data set is diverse and large enough, the observed proportion ought to be equivalent to the probability that applying the transformation to a new molecule (in the relevant context) will also cause the property to change in the same direction. This is often the kind of information that lead optimization projects are searching for, that transformations are likely to move a property in a desired direction or at the least not move it in an undesired direction. Similarly, data sets in which compounds are classified categorically, either because of the nature of the assay (e.g., Ames testing to identify either active or inactive compounds)[23] or artificially (e.g., Pfizer metabolism index),[24] can be analyzed in a pairwise fashion leading to predicted probabilities for changes among the various classifications. In their analysis of over 150 000 in vitro clearance measurements in human liver microsomes (HLM), Lewis and Cucurull-Sanchez quote 2-fold changes or greater as being significant based on variability in the experimental procedure.[25] Papadatos et al. also used a 2-fold margin in their classification of hERG activity, solubility, and lipophilicity into favorable, unfavorable, and zero effect transformations.[18] Despite loss of some quantitative information, this approach benefits from being able to include additional data points that are out of range of the assay. For example, a molecular transformation corresponding to a change in hERG activity from 7 to >30 μM could be considered as a significant decrease, albeit in the absence of any estimate of the magnitude of the reduction. In her study of Pfizer's uridine 5′-diphosphoglucuronosyltransferase (UGT) data set, Cucurull-Sanchez adopted a different approach by classifying the compounds themselves as exhibiting either high or low clearance before conducting any analysis.[26]

One complication with MMPA is deciding which properties are most useful for analysis. The simpler the property is, the more likely there is to be a straightforward link to molecular structure. This was exemplified by Leach et al.; effects upon solubility and plasma protein binding were shown to frequently occur with tight statistical certainty, whereas those concerning in vivo plasma exposure, which depends upon many other properties of the molecule, were less likely to do so.[4] Similarly, where a measurement relates simply to a binding event between protein and ligand, a better outcome might be expected than where a functional effect (agonism, antagonism, channel blocking, etc.) is measured, although we have found that MMPA can be effective for analyzing these more complex end points.

METHODS TO IDENTIFY MATCHED MOLECULAR PAIRS

This section focuses on the algorithmic details of how several groups have chosen to identify matched pairs, and those less computationally inclined might prefer to skip to the next section. Recent literature has demonstrated that there are a host of different approaches to identifying matched molecular pairs. The choices made by different researchers often provide an insight into the philosophy underpinning the use of matched molecular pairs in these different groups. In this section we highlight some of the key variations.

(i) Small Data Sets. There are numerous reports in the medicinal chemistry literature involving "matched pairs" within a series featuring a common core.[27−31] This basic form of MMPA is essentially an abstraction of the kind of information provided in many of the tables in publications in this journal where the variation of properties with substituents is detailed. It can be

formalized by defining a core and simply plotting the substituent at a given position against the property of interest.[32] The analysis is then completed with the connection of points in which the remainder of the molecule is the same. The magnitude and consistency of property differences between one substituent and another are revealed.[33] Automated MMPA in this form has been implemented at AstraZeneca to support lead optimization projects that work within defined series. A core is defined by each project team, and the various exchanges of substituents around the core are automatically detected and the corresponding changes in properties computed. This is particularly powerful in an environment in which all compounds are subject to parallel testing in a battery of assays such that a fairly comprehensive view of the effect structural changes have on the full range of properties can be detected; this in turn can drive multiparameter optimization.[34] Any new data that are generated are automatically added such that a simple Web tool can provide a project team with the most up to date view of the effect of all structural changes so far tried. This influences the choices that are made about which compound to make next in a dynamic fashion. In this simple form, the overlap with Free–Wilson analysis is clear. For the remainder of this section, only those methods where the formal definition of a core is not required are detailed.

(ii) Supervised Approaches. A number of attempts to correlate changes in molecular structure with changes in properties on a large scale have relied on the investigator, either wholly or in part, to define the transformations themselves prior to conducting the analysis. Researchers at AstraZeneca's Alderley Park site have used the molecular editor program LEATHERFACE to convert the SMILES for each putative member of a matched pair into its counterpart structure.[1,4,35,36] The use of SMARTS to define the substructure to be modified allows control over the chemical environment in which the structural modification occurs.[37] Further developments in the same group led to the program find_pairs and its successor thrice_pairs.[38,39] These rely upon defining the two end points with SMARTS patterns and defining how the atoms in each SMARTS pattern map onto one another and might be considered to be a basic implementation of SMIRKS.[4,11] Gleeson et al. used related methodology that also allows the second member of the pair to be encoded using SMARTS.[20] This extension allows less specific transformations such as ArCl to ArAr′ (Ar′ = any aromatic) to be studied and is achieved by using SMARTS filtering as opposed to SMILES matching as a final step. The diversity of the hit list can be expanded even further by using "X to any" type transformation definitions, where just the initial molecule less fragment X is used as a product substructure filter. Pfizer's "Buy me Grease" tool allows users to mine for matched molecular pairs via a Web interface.[25] The program is implemented in Pipeline Pilot and requires a RXN file to be provided by the user to define the molecular transformation of interest.[40] The reaction is applied to the entire data set and identifies compounds for which experimental data also exist for the products. Finally, a report is generated with a histogram indicating the percentage of times the transformation increased, decreased, or had no effect on the property, in vitro clearance in this case.

Other authors have used predefined transformation lists to supplement an unsupervised method that is reliant on the heavy atom framework of one member of the matched molecular pair being structurally embedded within the other. Hajduk and Sauer used the Daylight toolkit to identify instances where one member of each putative matched pair was a substructure of the other.[19,41] Cases where only one fragment remained following cleavage of that substructure were considered matched molecular pairs. Cucurull-Sanchez adopted a similar tactic, having identified a collection of substructural features that related to either high or low UGT clearance.[26]

(iii) Unsupervised Approaches. The fastest and most efficient way to identify matched molecular pairs within a data set in an unsupervised fashion is to decompose the molecules into fragments, which may then be indexed to allow rapid sorting and retrieval from a database. The fragmentation schemes are typically based on the method described by Bemis and Murcko, where acyclic single bonds are cleaved to define a molecular scaffold and its side chain substituents.[42,43] Such methods benefit from the requirement to process each molecule just once, as once the core(s) and side chain(s) for a given structure have been defined, the process of matched pair identification requires simply pairing compounds with a common core.

Haubertin and Bruneau used a variant of the RECAP fragmentation process in their analysis of over 50 000 druglike compounds to determine the effect of common structural transformations on solubility, plasma protein binding, and lipophilicity.[44,45] Their definition of a side chain was limited to fragments having an acyclic attachment point and a SMILES string of less than 40 characters, which resulted in a fixed library of 9038 substructures. Their rationale for limiting the size of the side chain library in this way was to expedite the identification of side chains in previously unprocessed structures submitted by users through a Web interface. All side chain occurrences were identified in each of the druglike molecules, generating a total of 386 757 core/side chain combinations, from which the core pairing process identified 733 445 occurrences of side chain substitution. Hussain and Rea were able to address some of the limitations of this approach by removing the 40-character limit from the definition of a fragment and extending the fragmentation procedure to include double and triple cut disconnection patterns.[46] Their analysis identified all heavy atom molecular pairs arising from acyclic bond disconnection from a set of 333 491 compounds in just under 14 h on a single CPU. Hydrogen substitutions were identified in a second step by capping the fragments arising from single cut fragmentations with hydrogen and then looking up the canonicalized SMILES among the input molecules. This paper reports the first use of SMIRKS to encode the transformations identified from MMPA, although the contextual information was not retained to ensure specificity in their application.

This issue of context dependence is one that has recently been studied in some depth at the University of Sheffield, U.K., in collaboration with GlaxoSmithKline.[18] The fragment indexing approach described above was selected for this study because of its computational efficiency. For each transformation, the context was characterized using a total of five additional descriptors, three describing the molecules as a whole (a reduced graph representation of the molecules, a Murcko scaffold, and a fingerprint based cluster membership), with the remaining two based on the local environment of the disconnection point (the nearest node from the reduced graph and a string representing the nearest three layers of SYBYL atom types).[42,47] Papadatos et al. reported a collection of common side chain replacements with well validated effects on hERG inhibition, solubility, and lipophilicity, most of which follow anticipated changes in lipophilicity. However, what makes this paper a must-read for anyone

interested in the field of MMPA are the examples where this expectation is defied in a way that can be attributed to the local environment of the attachment point and where there are enough examples to demonstrate that this is a real dependence. This paper offers a vision of how identifying context specific effects might be systematized.

A popular alternative to the acyclic disconnection algorithms described above are those that require a maximum common substructure search (MCSS) to be performed on each and every pair of molecules. The advantage of these algorithms is a conceptual one and relates back to the central definition of a matched molecular pair, i.e., two compounds that differ from each other by just a small well-defined structural change. As the name suggests, MCSS based approaches identify the maximum amount of substructure shared between two molecules; thus, only the absolute minimum substructure distinguishing the two structures is identified as a byproduct and forms the basis for molecular pairing.

Unfortunately, this method of identifying matched molecular pairs comes at price, as where the fragmentation algorithms require each molecule to be processed only once, a complete MCSS based approach requires a full $n \times n$ comparison of all the molecules in a data set. To overcome the computational expense incurred, Sheridan performed clustering on his data set of 100 000 compounds from the MDL Drug Data Report (MDDR), first based on biological activity and then on topological descriptors prior to performing MMPA by MCSS.[48,49] All pairs of compounds within a cluster were compared with each other, with the highest scoring common substructure (HSCS) being used as a core from which the fragments undergoing modification could be identified.[50] Having ensured that a single fragment was all that set the two molecules apart, the remainder of the molecule was pruned according to two methods, one which encoded just the bare minimum number of atoms required to define the transformation (algorithm A) and one where rings directly encompassing the transformation were also preserved (algorithm B). As the compounds had been clustered according to their biological activity at the outset, the most frequently occurring transformations were considered bioisosteres, as they represented groups that could be interchanged with each other, retaining activity across a variety of target classes. Sheridan reports that despite returning larger fragments and fewer observations, algorithm B appeared to be "more appealing because it retains more of the contextual information of the replacement".

A revised version of Merck's algorithm, published in 2006, went beyond simply identifying molecular pairs with activity against a given receptor and began to associate quantitative changes in activity between the pair members.[5] Dopamine 2 (D2) agonism, dihydrofolate reductase (DHFR) inhibition, and angiotensin converting enzyme inhibition were studied. First, the T-ANALYZE program was used to identify the "remainder after elimination of common substructure" (RECS) for pairs of molecules satisfying a given degree of similarity, which defined the nature of the transformation itself in the previously reported "algorithm A". Examples of each transformation were then clustered according to atom-pair and topological-torsion descriptors to ensure that the transformation occurred within a similar context in all cases, and the experimental data were added to each of the clusters as a final step. The clusters were scored and ranked allowing users to easily visualize interesting regions of consistent, local SAR. The authors then took the D2 and DHFR data sets and submitted a probe molecule to a second program, T-MORPH, which identified relevant transformations that could be applied to the probe and their likely effects on activity. In AstraZeneca's most recently published method, the authors also use a MCSS to identify molecular pairs, only having identified the RECS, they proceed to incorporate a series of layers of common substructure in a hierarchical version of Merck's "algorithm B".[51] The advantage of this so-called "WizePairZ" approach is that the local environment is encoded as part of the transformation (using nonproprietary SMIRKS notation). Having identified transformations and effects on histone deacetylase inhibition from a small collection of compounds, all were fed back into the program, resulting in the generation of two structures initially omitted from the data set and accurate predictions of their biological activity.

Improvements in computer performance over recent years have seen ever larger data sets mined for molecular transformations by MCSS. While retaining a topological similarity based cutoff, Southall and Ajay were able to do away with biological target-based clustering in their analysis of 116 550 kinase inhibitors from patents and the medicinal chemistry literature.[52] They estimated that by setting the Tanimoto threshold for comparison at 0.5, they were able to capture 90% of the molecular pairs in the data set while reducing the number of MCSS searches by 1000-fold. The calculation took approximately 2 weeks on a 30-processor cluster. Warner et al. have also reported similar performance, performing an entirely unfiltered analysis of over 35 000 compounds in around 3 days on 36 CPUs.[51] In what is undoubtedly the largest such study to date, researchers at Eli Lilly performed around 3.6×10^{12} MCSS calculations in their analysis of almost 2.7 million structures.[53] Without clustering based on either biological activity or molecular similarity, the time taken to complete this enormous calculation was reduced through use of the RASCAL algorithm for the MCSS,[54] the use of constraints on the extent of common substructure required and shortest path lengths between connection points, the omission of hydrogen replacements, and the use of a 1072 core cluster on which to perform the comparisons.[53−55] Nevertheless this study marks a significant achievement, as 2.7 million structures are many more than are ever likely to exist in an experimental data set to be mined by MMPA.

FUTURE FOR MATCHED MOLECULAR PAIRS

Having outlined some of the principal considerations in how and when to apply MMPA, it is worth considering what new avenues might be followed. The first priority is to apply this kind of analysis to larger and more diverse sets of compounds, now that the feasibility of such comparisons has been demonstrated,[53] and to data sets generated in new assay types not yet considered. This ought to lead to new insights and guidelines of utility to understanding and optimizing compounds. During the course or preparing this manuscript, a number of pertinent publications have appeared: Hu and Bajorath have found transformations that change a compound's biological activity profile (with 754 different target proteins considered); Wassermann and Bajorath identified bioisosteric changes that are target family specific; Keefer et al. studied effects on a range of ADME properties caused by a diverse set of structural changes.[2,56,57] As more parameters are studied, the possibility of identifying transformations that facilitate multiparameter optimization should be feasible. By way of illustration of this and the predictive power of even the so far limited published analyses, we note that in a

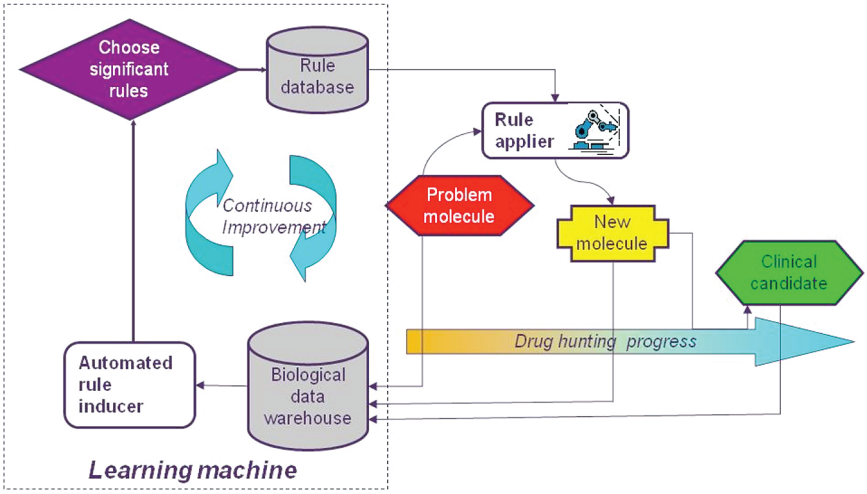

Figure 7. Flow diagram for a potential automated drug design tool. In the left-hand box, databases are mined to obtain rules; matched molecular pair analysis provides a simple and effective way to obtain these rules. In the right-hand section, these rules are applied to compounds that have deficiencies preventing them from becoming clinical candidates to suggest new molecules that might overcome those deficiencies.

recent publication from AstraZeneca it is reported that addition of a *m*-Me group in a particular compound reduces solubility from 9.5 to 8 μM and increases Cl_{int} (in microsomes) from 76 to 100 μL min^{-1} mg^{-1}.[58] These changes correspond to log(Sol) and log(Cl_{int}) both decreasing by 0.1 which agrees well with expected values of 0.21 and 0.28 predicted by Leach et al. and Dossetter, respectively.[4,11] We hope that publications concerning a broader range of structural changes relating to more properties will lead to more examples where MMPA has predicted structural changes that have changed a number of properties in such a way as to strike a desired balance.

One limitation of the approach, as stressed above, is that it can only make predictions about structural features that have precedent in any given assay. This might be circumvented by introducing levels of abstraction for particular groups; rather than grouping transformations in a way that requires an exact match, this could be loosened. For instance, if a data set does not contain any matched pairs in which an ethyl sulfone has been added to an aromatic ring but does have a set of pairs in which a methyl sulfone has been added, then the prediction might be based on this with a lipophilicity (or alternative substituent related descriptor such as σ) term added to compensate for the larger alkyl group. The challenge will be to ensure that the user is comfortable as the analysis moves away from being a means of organizing and presenting experimental data. The concepts being developed in the area of molecular graph theory describing the amount of editing required to link two structures may contribute to the organizing of transformations both to address the issues highlighted in Figure 4 and to facilitate the abstraction necessary for extrapolation.[59]

The pairwise comparison involved in MMPA ought to complement and validate computational modeling methods that either rely upon abbreviating to model systems (as is the case in many quantum mechanical studies) or that have many errors involved that might otherwise mask any signal (as is the case in many docking/scoring or even simulation studies). In the former case, expensive calculations on protein−ligand binding are only currently feasible when focused on small regions of the protein that are in proximity to the ligand, analogous to the theozymes proposed by Houk and co-workers to rationalize enzymatic catalysis.[60] While calculation of individual binding free energies will be impossible, computed changes in binding energy for structural changes ought to relate to the sort of differences that MMPA produces. Equally, the ab initio calculation of changes in intrinsic solubility from lattice and solvation energies could be validated with MMPA.[61] By contrast, modeling all of the atoms in a complex system requires a number of approximations to be used (such as those in pose scoring or in force fields) that reduce accuracy. Computing differences could benefit from cancellation of some of the errors introduced by these approximations, and these differences can be compared to the output of MMPA. The output of MMPA is limited to the observation that a structural change leads to a particular property change but often prompts the question of why this might be; these two approaches can help to address this.

A view of how matched molecular pairs and other analyses might contribute most effectively to pharmaceutical research has been described by Griffen, outlined in Figure 7.[62] Here, all available and relevant databases can be examined to find rules that link changes in chemical structure with improvements in properties of medicinal chemistry interest; this happens in the dashed box to the left of Figure 7. These rules can then be applied to "problem molecules" that have one or more properties preventing them from becoming clinical candidates. The application of the rules is performed in the right-hand section of Figure 7. In principle, the identification and application of the rules do not need manual intervention.

MMPA provides a simple way to explore data sets in search of the rules required for this kind of approach. The various unsupervised approaches described above can all be used to do just this. Each will mine a data set and find sets of compounds that differ only by small, well-defined transformations. They can also compute for each the corresponding effect on a property of interest. When such methods are applied to data sets of the size typically available (thousands of compounds), they will discover enormous sets of potential rules (thousands to millions).[51] These rules are subject to the issues highlighted in Figure 4 of differing degrees of specificity contrasting with different degrees of exemplification. In the context of generating rules, these issues translate to the dilemma of whether high degrees of certainty for

rules derived from many diverse molecules are preferred over lower degrees of certainty for larger, structurally more specific effects. The choice will depend upon the philosophy of those involved and is unlikely to have any absolute right answer.

One consequence of taking an approach like that outlined in Figure 7 is that the nature of the compounds that are represented in the databases of measurements becomes highly significant. Ideally, significant numbers of structural transformations would be represented with each deriving from a diverse and large set of compounds. Most of the measurements are generated without this consideration; they are generated to determine properties of interest to particular drug discovery projects. By identifying compounds that would populate matched pairs sets that are available for testing and generating data on them, MMPA can improve its own utility by enhancing the certainty of rules that appear to be interesting but that do not have enough data to be convincing. In this way, the ideal outcome of providing drug designers with rules that surprise them but that are nonetheless convincing can be achieved.

SUMMARY

MMPA has become accepted as a useful approach in drug discovery. In the absence of experimental data, it can be used to identify the frequency with which certain transformations have been performed. It has been applied across a range of properties of medicinal chemistry interest. It contrasts favorably with some alternative data analysis methods and in particular presents experimental data to drug designers in a way that naturally suggests what compound to make next but within the limit of transformations that have been exemplified in the data set. It can be used to define a set of transformations that have historically modulated a given property that can assist the design of future compounds. Many researchers have developed software for identifying matched molecular pairs, and the approach taken tends to be governed by the motivation behind the analysis. There are a number of important considerations determining when matched pairs are appropriate. There is an inevitable trade-off between structurally specific changes that are less well exemplified and more general changes that are better represented but less likely to be surprising. Data ought to be carefully prepared and generated in assays that are as homogeneous as is practicable. Matched pairs can complement molecular modeling techniques such as quantum mechanics and docking. Matched pairs can be used to obtain rules of thumb that can guide improvements in any measurable property or indeed several properties at once.

AUTHOR INFORMATION

Corresponding Author

*Phone: +44 1625 231853. E-mail: andrew.leach@astrazeneca.com.

BIOGRAPHIES

Ed Griffen is a Principal Scientist working in the Oncology Innovative Medicines unit at AstraZeneca. He studied for his Ph. D. at Imperial College, London, in marine alkaloid synthesis with Professors Charles Rees and Christopher Moody and then joined the research group of Professor Victor Snieckus as a Postdoctoral Fellow developing methods in indole metalation and radiometal chelation chemistries. He is a named inventor on 16 patents and is a coauthor of the textbook "On Medicinal Chemistry". Ed is a Visiting Lecturer at the University of Manchester, U.K., in the medicinal chemistry program. Recent research publications have been in the application of data mining methods to accelerate drug discovery.

Andrew G. Leach is a computational chemist working in the Cardiovascular and Gastrointestinal Innovative Medicines unit at AstraZeneca. He obtained his Ph.D. in the group of Professor S. V. Ley at the University of Cambridge, U.K., for work developing new techniques for combinatorial chemistry. He undertook postdoctoral studies at the University of California, Los Angeles, in the group of Professor K. N. Houk, investigating several pericyclic reaction mechanisms and the interactions between proteins and transition states.

Graeme R. Robb is a computational chemist working in the Cardiovascular and Gastrointestinal Innovative Medicines Unit at AstraZeneca. He obtained his Ph.D. in chemical physics under the supervision of Professor A. Harrison, studying the effects of microwave radiation on solid materials using in situ X-ray diffraction. He joined AstraZeneca in 2002 and has worked predominantly in lead optimization programs, with particular interest in GPCR receptors.

Daniel J. Warner is a computational chemist with a background in biomolecular simulation and analysis. He completed his Ph.D. under the supervision of Drs. Charles Laughton and Stephen Doughty at the University of Nottingham, U.K., before joining AstraZeneca in October 2006. Since then, he has supported predominantly structure-based lead optimization programs, first within respiratory and inflammation at Charnwood in the U.K. and more recently in neuroscience at AstraZeneca R&D Montreal, Canada. He is currently involved in the delivery of a global system for improved exploitation of MMPA across AstraZeneca.

DEDICATION

†This article is dedicated to Dr. Andy Barker on the occasion of his recognition by the American Chemical Society as a Hero of Chemistry and of his retirement from AstraZeneca.

ABBREVIATIONS USED

MMPA, matched molecular pair analysis; SAR, structure−activity relationship; SPR, structure−property relationship; QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; EGFR, epidermal growth factor receptor tyrosine kinase; RMSE, root-mean-square error; SEM, standard error in the mean; SD, standard deviation; MCSS, maximum common substructure; hERG, human ether-a-go-go-related gene encoded potassium channel; HLM, human liver microsomes; RECS, remainder after elimination of common substructure

REFERENCES

(1) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. *Methods Princ. Med. Chem.* **2005**, *23*, 271–285.

(2) Keefer, C. E.; Chang, G.; Kauffman, G. W. Extraction of tacit knowledge from large ADME data sets via pairwise analysis. *Bioorg. Med. Chem.* **2011**, *19*, 3739–3749.

(3) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195–198.

(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) 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.

(6) Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **2011**, *54*, 2529–2591.

(7) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Mol. Inf.* **2010**, *29*, 366–385.

(8) Free, S. M., Jr.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395–399.

(9) Patel, Y.; Gillet, V. J.; Howe, T.; Pastor, J.; Oyarzabal, J.; Willett, P. Assessment of additive/nonadditive effects in structure−activity relationships: implications for iterative drug design. *J. Med. Chem.* **2008**, *51*, 7552–7562.

(10) Topliss, J. G. A manual method for applying the hansch approach to drug design. *J. Med. Chem.* **1977**, *20*, 463–469.

(11) Dossetter, A. G. A statistical analysis of in vitro human microsomal metabolic stability of small phenyl group substituents, leading to improved design sets for parallel SAR exploration of a chemical series. *Bioorg. Med. Chem.* **2010**, *18*, 4405–4414.

(12) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure−activity relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.

(13) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev. (Washington, DC, U. S.)* **2010**, *110*, 5714–5789.

(14) Guha, R.; Van Drie, J. H. Assessing how well modeling protocol captures structure−activity landscape. *J. Chem. Inf. Model* **2008**, *48*, 1716–1728.

(15) Livingstone, D. J.; Salt, D. W. Judging the significance of multiple linear regression models. *J. Med. Chem.* **2005**, *48*, 661–663.

(16) Bruneau, P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.

(17) *JMP*, version 8.0; SAS Institute Inc.: Cary, NC.

(18) 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.

(19) Hajduk, P. J.; Sauer, D. R. Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* **2008**, *51*, 553–564.

(20) Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. ADMET rules of thumb II: a comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem.* **2009**, *17*, 5906–5919.

(21) Fujita, T.; Ban, T. Structure−activity relation. 3. Structure−activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. *J. Med. Chem.* **1971**, *14*, 148–152.

(22) Wassermann, A. M.; Bajorath, J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

(23) Ames, B. N.; Gurney, E. G.; Miller, J. A.; Bartsch, H. Carcinogens as frameshift mutagens. Metabolites and derivatives of 2-acetylaminofluorene and other aromatic amine carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* **1972**, *69*, 3128–3132.

(24) Lewis, R. A. Computer-aided drug design 2005−2007. *Chem. Modell.* **2008**, *5*, 51–66.

(25) Lewis, M. L.; Cucurull-Sanchez, L. Structural pairwise comparisons of HLM stability of phenyl derivatives: introduction of the Pfizer metabolism index (PMI) and metabolism-lipophilicity efficiency (MLE). *J. Comput.-Aided Mol. Des.* **2009**, *23*, 97–103.

(26) Cucurull-Sanchez, L. Successful identification of key chemical structure modifications that lead to improved ADME profiles. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 449–458.

(27) Gall, M.; McCall, J. M.; TenBrink, R. E.; VonVoigtlander, P. F.; Mohrland, J. S. Arylformamidines with antinociceptive properties. *J. Med. Chem.* **1988**, *31*, 1816–1820.

(28) Blaszczak, L. C.; Brown, R. F.; Cook, G. K.; Hornback, W. J.; Indelicato, J. M.; Jordan, C. L.; Katner, A. S.; Kinnick, M. D.; McDonald, J. H., III; Morin, J. M.; Munroe, J. E.; Pasini, C. E. Comparative reactivity of 1-carba-1-dethiacephalosporins with cephalosporins. *J. Med. Chem.* **1990**, *33*, 1656–1662.

(29) Herberich, B.; Cao, G.; Chakrabarti, P. P.; Falsey, J. R.; Pettus, L.; Rzasa, R. M.; Reed, A. B.; Reichelt, A.; Sham, K.; Thaman, M.; Wurz, R. P.; Xu, S.; Zhang, D.; Hsieh, F.; Lee, M. R.; Syed, R.; Li, V.; Grosfeld, D.; Plant, M. H.; Henkle, B.; Sherman, L.; Middleton, S.; Wong, L. M.; Tasker, A. S. Discovery of highly selective and potent p38 inhibitors based on a phthalazine scaffold. *J. Med. Chem.* **2008**, *51*, 6271–6279.

(30) Shi, Y.; Sitkoff, D.; Zhang, J.; Klei, H. E.; Kish, K.; Liu, E. C. -.; Hartl, K. S.; Seiler, S. M.; Chang, M.; Huang, C.; Youssef, S.; Steinbacher, T. E.; Schumacher, W. A.; Grazier, N.; Pudzianowski, A.; Apedo, A.; Discenza, L.; Yanchunas, J.; Stein, P. D.; Atwal, K. S. Design, structure-activity relationships, X-ray crystal structure, and energetic contributions of a critical P1 pharmacophore: 3-chloroindole-7-yl-based factor Xa inhibitors. *J. Med. Chem.* **2008**, *51*, 7541–7551.

(31) Ioannidis, S.; Lamb, M. L.; Wang, T.; Almeida, L.; Block, M. H.; Davies, A. M.; Peng, B.; Su, M.; Zhang, H.; Hoffmann, E.; Rivard, C.; Green, I.; Howard, T.; Pollard, H.; Read, J.; Alimzhanov, M.; Bebernitz, G.; Bell, K.; Ye, M.; Huszar, D.; Zinda, M. Discovery of 5-chloro-N2-[(1*S*)-1-(5-fluoropyrimidin-2-yl)ethyl]-N4-(5-methyl-1*H*-pyrazol-3-yl)pyrimidine-2,4-diamine (AZD1480) as a novel inhibitor of the Jak/Stat pathway. *J. Med. Chem.* **2011**, *54*, 262–276.

(32) Andrews, D. M.; Gibson, K. M.; Graham, M. A.; Matusiak, Z. S.; Roberts, C. A.; Stokes, E. S. E.; Brady, M. C.; Chresta, C. M. Design and campaign synthesis of pyridine-based histone deacetylase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 2525–2529.

(33) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(34) Andersson, S.; Armstrong, A.; Bjoere, A.; Bowker, S.; Chapman, S.; Davies, R.; Donald, C.; Egner, B.; Elebring, T.; Holmqvist, S.; Inghardt, T.; Johannesson, P.; Johansson, M.; Johnstone, C.; Kemmitt, P.; Kihlberg, J.; Korsgren, P.; Lemurell, M.; Moore, J.; Pettersson, J. A.; Pointon, H.; Ponten, F.; Schofield, P.; Selmi, N.; Whittamore, P. Making medicinal chemistry more effective—application of Lean Sigma to improve processes, speed and quality. *Drug Discovery Today* **2009**, *14*, 598–604.

(35) Birch, A. M.; Kenny, P. W.; Simpson, I.; Whittamore, P. R. O. Matched molecular pair analysis of activity and properties of glycogen phosphorylase inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 850–853.

(36) Morley, A. D.; Kenny, P. W.; Burton, B.; Heald, R. A.; MacFaul, P. A.; Mullett, J.; Page, K.; Porres, S. S.; Ribeiro, L. R.; Smith, P.; Ward, S.; Wilkinson, T. J. 5-Aminopyrimidin-2-ylnitriles as cathepsin K inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 1658–1661.

(37) (a) *SMARTS*; Daylight Chemical Information Systems Inc.: Santa Fe, NM, Vol 471. (b) http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, .

(38) Jones, H. D.; Cosgrove, D. A. *find_pairs*; AstraZeneca Pharmaceuticals: Macclesfield, U.K.

(39) Cosgrove, D. A. *thrice_pairs*; AstraZeneca Pharmaceuticals: Macclesfield, U.K.

(40) *Pipeline Pilot*, version 6.1.5; Accelrys, Inc.: San Diego, CA.

(41) *Daylight*, version 4.83; Daylight Chemical Information Systems, Inc.: Laguna Niguel, CA.

(42) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(43) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.

(44) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(45) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
(46) 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.
(47) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(48) *MDDR*, version 1; Accelrys, Inc.: San Diego, CA.
(49) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(50) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 915–924.
(51) 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.
(52) Southall, N. T. Ajay kinase patent space visualization using chemical replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.
(53) Raymond, J. W.; Watson, I. A.; Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.* **2009**, *49*, 1952–1962.
(54) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305–316.
(55) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–533.
(56) Hu, Y.; Bajorath, J. Chemical transformations that yield compounds with distinct activity profiles. *ACS Med. Chem. Lett.* **2011**, *2*, 523–527.
(57) Wassermann, A. M.; Bajorath, J. Identification of target family directed bioisosteric replacements. *MedChemComm* **2011**, *2*, 601–606.
(58) Alhambra, C.; Becker, C.; Blake, T.; Chang, A.; Damewood, J. R., Jr.; Daniels, T.; Dembofsky, B. T.; Gurley, D. A.; Hall, J. E.; Herzog, K. J.; Horchler, C. L.; Ohnmacht, C. J.; Schmiesing, R. J.; Dudley, A.; Ribadeneira, M. D.; Knappenberger, K. S.; Maciag, C.; Stein, M. M.; Chopra, M.; Liu, X. F.; Christian, E. P.; Arriza, J. L.; Chapdelaine, M. J. Development and SAR of functionally selective allosteric modulators of GABAA receptors. *Bioorg. Med. Chem.* **2011**, *19*, 2927–2938.
(59) Heinonen, M.; Lappalainen, S.; Mielikainen, T.; Rousu, J. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J. Comput. Biol.* **2011**, *18*, 43–58.
(60) Tantillo, D. J.; Chen, J.; Houk, K. N. Theozymes and compuzymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* **1998**, *2*, 743–750.
(61) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharmaceutics* **2008**, *5*, 266–279.
(62) Griffen, E. The rise of the intelligent machines in drug hunting? *Future Med. Chem.* **2009**, *1*, 405–408.