Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity

George Papadatos, Muhammad Alkarouri, Valerie J. Gillet,* and Peter Willett

Information School, University of Sheffield, Sheffield S1 4DP, U.K.

Visakan Kadirkamanathan

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K.

Christopher N. Luscombe, Gianpaolo Bravi, Nicola J. Richmond, Stephen D. Pickett, Jameed Hussain, John M. Pritchard, Anthony W. J. Cooper, and Simon J. F. Macdonald

GlaxoSmithKline Medicines Research Centre, Stevenage SG1 2NY, U.K.

Received July 8, 2010

Previous studies of the analysis of molecular matched pairs (MMPs) have often assumed that the effect of a substructural transformation on a molecular property is independent of the context (i.e., the local structural environment in which that transformation occurs). Experiments with large sets of hERG, solubility, and lipophilicity data demonstrate that the inclusion of contextual information can enhance the predictive power of MMP analyses, with significant trends (both positive and negative) being identified that are not apparent when using conventional, context-independent approaches.

INTRODUCTION

Lead optimization is a complex, time-consuming task, in which the medicinal chemist seeks to obtain a sufficiently promising balance among potency, off-target interactions, toxicity, and pharmacokinetic behavior, inter alia, to make it worth allowing a molecule to progress to the candidate stage of the drug discovery pipeline. The successful optimization of an initial lead compound is hence crucially dependent on the medicinal chemist's ability to choose which analogue (or set of analogues when, as is very often the case, chemical arrays are being used) should be synthesized next, based on the knowledge that has been obtained thus far in the optimization. For example, there might be a need for analogues that are notably more soluble but, as much as possible, still have the same (or even higher) potency.

Many chemoinformatic approaches are available to assist in lead optimization.[1] For example, approaches based on bioisosterism suggest substructural transformations that can be applied to an existing bioactive lead to yield analogues that might exhibit superior pharmacokinetic properties.[2−6] Quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) approaches have been used for optimization for many years[7−10] and express the property of interest as a function of structural variables characterizing the molecules in a training set (typically those that have been synthesized and tested thus far in the optimization). Given such an expression, property values can then be computed for previously untested analogues, thus enabling the chemist to predict the change in property, ΔP, resulting from a specific structural change, ΔS. This is a powerful technique and one that is widely used; however, the medicinal chemist is likely to be at least as interested in the reverse process of deriving ΔS given a desired ΔP value. The ability to predict the change in structure that is required to bring about a specific change in property, or inverse QSAR,[11,12] is the basis for the work reported here, which belongs to the class of matched molecular pair (MMP) methods that have come to the fore over the past few years.[12−18] These methods are related to the bioisosterism approaches mentioned above in their focus on specific substructural transformations; however, they go further in providing quantitative estimates of the changes, ΔP, that result from the application of particular transformations, ΔS, and hence provide an inverse QSAR approach to optimization. Another difference is that, given an appropriate source of data, they can model not only biological activity, the principal focus of the bioisosterism approaches, but also any chemical, physicochemical, or ADME (absorption, distribution, metabolism, and excretion) property that needs to be optimized.

An MMP is defined as two molecules that differ from each other by a small, specified change at one or more specified locations and that share a large, identical structural feature. We refer to the change between the pair as a transformation and the invariant feature as the context, with the point where the transformation has taken place referred to as the attachment point. These terms are illustrated in Figure 1, which shows an example of a single-point transformation, where there is only one attachment point. Multiple-point transformations are also possible, but they have not been considered in the current study. Efficient algorithms for MMP identification have been described by Raymond et al.[19] and by Hussain and Rea.[20]

* Corresponding author e-mail: v.gillet@sheffield.ac.uk; phone: 0044-114-2222652.

10.1021/ci100258p © 2010 American Chemical Society
Published on Web 09/28/2010

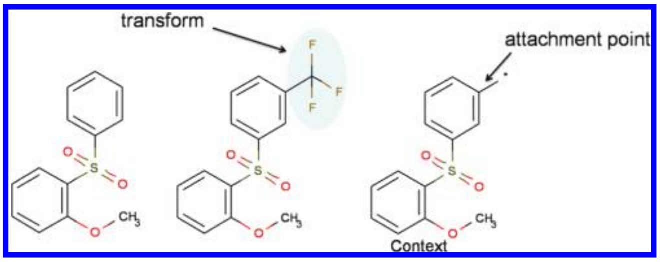


Figure 1. One matched molecular pair and its context. The transformation is H → CF_3 (a single-point change) and is highlighted in yellow. The asterisk in the context denotes the attachment point.

An early MMP study by Sheridan et al. identified molecular transformations using a maximum common substructure (MCS) procedure that was applied to a subset of the MDL Drug Data Report (MDDR) database with potency as the end point.[13] This is the only MMP study that has used public, as opposed to corporate, data as the basis for analysis. Hajduk and Sauer at Abbott Laboratories analyzed additions (i.e., H → Y), functional group transformations (e.g., Br → OMe), and multiple regiospecific phenyl substitutions and noted that the potency changes caused by most transformations followed a normal distribution centered near zero.[14] Leach et al. at AstraZeneca focused on ADME properties, such as aqueous solubility, plasma protein binding, and oral exposure, and molecular transformations, such as phenyl ring additions (i.e., Ph−H → Ph−Y) and methylation of heteroatoms (e.g., OH → OCH_3).[21] Haubertin and Bruneau, also at AstraZeneca, analyzed a set of ca. 9000 functional groups and their effect on solubility, protein binding, and distribution coefficient (log *D*), as well as several computed physicochemical properties.[12] Lewis and Cucurull-Sanchez at Pfizer investigated the effects of regiospecific single and double phenyl ring additions using in-house data on human liver microsome activity and intrinsic clearance.[15] Perhaps the most comprehensive study is that of Gleeson et al. at GlaxoSmithKline, who reported a systematic analysis of ca. 500 000 ADMET (absorption, distribution, metabolism, elimination, and toxicity) data points generated in eight in vitro assays for P450 inhibition, hERG (human ether-à-go-go-related gene) inhibition, solubility, and permeability.[16] The study considered only additions (H → Y) where Y was a member of a predefined list of ca. 90 frequently used substituents. Finally, in two practical applications of the MMP method, Birch et al. described the design of inhibitors of glycogen phosphorylase,[17] and Southall and Ajay described the analysis of protein kinase patents.[18]

An assumption of the basic MMP approach is that the property difference, Δ*P*, resulting from a specific transformation depends only on the substructural change that has taken place, irrespective of the context, that is, of the structural environment in which that transformation has taken place. For example, if a hydrogen atom is replaced by a trifluoromethyl group (H → CF_3), then the change in lipophilicity that results is assumed to be constant across the many potential contexts in which that transformation might take place. This is clearly a very strong assumption, and some of the published MMP studies have hence made limited use of contextual information to enhance the specificity of the analyses that can be carried out. For example, Gleeson et al. differentiated halogen, amine, and alcohol additions depending on whether aromatic or aliphatic addition was involved.[16]

The starting point for the work reported here is a belief that the power of the MMP approach can be substantially increased if full use is made of the contextual information that is available when data mining is carried out on a sufficiently large scale (as is possible with the corporate data archives of the major pharmaceutical companies). We hence consider the context as an inherent component of an MMP analysis, and one that can be used to enhance the practical utility of the relationships existing between Δ*P* and Δ*S*. A further distinguishing characteristic of our work is that the MMPs are identified using a completely unsupervised process. Thus, as advocated by Sheridan et al.[13] and by Hussain and Rea,[20] there are no predefined lists of substituents or transformations, and we are hence not restricted to, for example, particular functional groups or addition reactions. In this article, we describe in detail the context-based approach to MMP analyses that we have developed and the application of our method to large sets of hERG, solubility, and lipophilicity data.

METHODS

Data Sets. The work reported here uses structural and property data from the GlaxoSmithKline (GSK) corporate database relating to three important ADME properties: hERG inhibition, solubility, and lipophilicity.

Human ether-à-go-go-related gene (hERG) codes for the homonymous potassium ion channel protein. Inhibition of the hERG ion channel is an important antitarget in drug discovery as it is associated with potentially fatal heart conditions, and compounds are routinely assayed against it during lead optimization.[22] A GSK fluorescence polarization (FP) in vitro assay was used to obtain the hERG inhibition measurements studied here. This assay relies on the binding of a fluorescent ligand to hERG membranes, as potential hERG ligands compete with the fluorescent compound and cause a decrease in FP signal. Solubility is one of the most important physicochemical properties in drug discovery: Low-solubility compounds suffer from poor absorption and bioavailability after oral dosing, and they can also cause synthetic and developmental problems.[23] The solubility measurements here came from a GSK chemiluminescent nitrogen detector (CLND) solubility assay. After the preparation of the sample from 10 mM dimethyl sulfoxide (DMSO) stock solution, the response from the detector is directly proportional to the number of nitrogen atoms in the sample; hence, by knowing this number beforehand, it is easy to determine the molecule's concentration in solution.[24] Finally, lipophilicity is another physicochemical property of major importance that directly affects both biological activity and ADMET properties.[23] The lipophilicity measurements came from a GSK chromatographic log *D* assay that measures a molecule's gradient retention time in reverse-phase high-performance liquid chromatography (HPLC) at a pH of 7.4.

For each of the three data sets, measurements containing modifier values (i.e., reported as ">" or "<") were filtered out, with the exception of the solubility data set where the values with modifiers were kept. The rationale behind this approach is that such modifiers indicate an insoluble or very soluble compound, thus providing valuable information that should not be omitted. Because the statistics employed here consider bins of data and not the actual ΔP values, the decision to include the modified values does not interfere with the reliability of the findings. Furthermore, multiple end-point values for a single compound were averaged. The Simplified Molecular Input Line Entry Specifications (SMILES) of the molecules were processed as follows: Stereochemistry information was removed, any charges were standardized, additional fragments and salts were removed, the SMILES strings were canonicalized, and duplicate or invalid structures removed. Although it is acknowledged that ignoring stereochemistry could lead to invalid matched pairs (such as a match between R- and S- isomers), the incidence of such pairs is likely to be minimal compared to the total number of MMPs. Moreover, when dealing with fragments such as substituents or contexts, the stereochemistry can change depending on the groups that are removed during fragmentation. The resulting data sets contained 76 266, 94 053, and 180 440 molecules for which hERG, solubility, and lipophilicity measurements, respectively, were available. The means and standard deviations of the property value for these three data sets were as follows: hERG measured by pIC50, 5.44 ± 0.74; solubility measured by log(micromoles per liter), 1.84 ± 0.71; and lipophilicity measured by log D, 4.55 ± 1.95.

MMP Identification. Our initial studies focused on the hERG data set, for which two computational procedures were developed to identify the MMPs present.

The identification of the MMPs in the first approach uses the dt_commonsubstruct and findsubs functions contained in the Daylight Toolkit v4.94 (available from Daylight Chemical Information Services, Inc., at http://www.daylight.com). The first function identifies and returns the MCS between a pair of molecules. Next, using the second function, the MCS found is mapped back onto the two original molecules, and the unmatched parts of these molecules then comprise the transformation. The algorithm can identify matched pairs exhibiting both single- and multiple-point transforms. In principle, the algorithm needs to be invoked $N(N - 1)/2$ times for a data set of N molecules; in practice, pairs of molecules were considered for MCS matching only if they had a Daylight Tanimoto similarity greater than 0.75 and differed by ≤12 in their heavy-atom counts.

The identification of the MMPs in the second approach uses the algorithm described recently by Hussain and Rea.[20] This works by fragmenting each molecule and then storing and indexing all of the enumerated fragments in an inverted-file-like structure. For each pair of molecules, the algorithm returns all of the fragments that the pair has in common and, thus, all possible matched pairs. The algorithm can identify only matched pairs exhibiting single transformations. However, it can process pairs having disconnected common substructures (i.e., where the transformation takes place in the middle of the molecular structure), although we excluded all such transformations here and considered only terminal substructural changes. Pairs of molecules are processed if they differ by ≤15 in their heavy-atom counts; moreover, after appropriate modification of the algorithm, the shared, unchanged part of an MMP must be at least as large (in terms of numbers of heavy atoms) as the change itself.

Table 1. Types and Occurrences of Transformations Identified in the hERG Data Set by the Two MMP Algorithms

transformation	Daylight MCS	fragment indexing
H → CH_3	8382	8326
H → F	4271	4243
H → Cl	2872	2849
H → OCH_3	2488	2484
F → Cl	1435	1441
F → OCH_3	1103	1120
CH_3 → Cl	1103	1106
H → CF_3	1085	1094
CH_3 → F	1070	1071
CH_3 → OCH_3	1012	1017
H → Et	920	986
Cl → OCH_3	913	927

The first algorithm is a "pure" MCS algorithm that seeks to match each and every heavy atom and bond in the pair of molecules, whereas the fragmentation procedure at the heart of the second, which is henceforth called the fragment indexing algorithm, fragments only acyclic single bonds between two heavy atoms. It is easy to find cases where the two algorithms will yield different results; in general, however, they are of comparable effectiveness. For example, Table 1 lists the numbers of occurrences for the 12 most common transformations identified in the hERG data set (which contains over 2.9×10^9 distinct pairs of molecules). The level of agreement in the table is striking; however, the computational requirements of the two algorithms are very different: The Daylight MCS algorithm required ca. 4.5 days on a four-processor machine after application of the initial Tanimoto filtering, whereas the fragment indexing algorithm required ca. 8 h on the same machine without such filtering. In light of this comparison, the latter algorithm was used for all of the experiments reported here.

Characterization of the Context. We used both global and local descriptors to characterize the contexts of MMPs: A global descriptor encodes an entire molecule other than the features involved in the transformation, whereas a local descriptor encodes the structure in the immediate vicinity of the attachment point.

There were three global descriptors: reduced graphs, Murcko frameworks, and Daylight fingerprints. A reduced graph (RG) is an abstraction of a molecule that uses graph nodes to represent separate structural and pharmacophoric features such as rings, linkers, positive/negative ionizability, and hydrogen-bond donor/acceptor groups. A molecule is thus represented by a set of connected nodes analogous to the atoms that comprise the nodes of a conventional chemical graph. For the purposes of this study, in-house software was used that generates RGs encoded by a modified SMILES string in which each RG feature is represented by the elemental symbol of an uncommon atom type.[25] Murcko frameworks provide a less abstract molecular representation, in which the side chains of a structure are trimmed, preserving only the rings and the linkers that connect these rings.[26] An in-house script was used to generate customized Murcko frameworks, in which atom and bond types are preserved, as is the attachment point, whereas terminal side chains and three- to eight-membered aliphatic rings such as cyclopropyl and cyclobutyl rings are trimmed. The resulting Murcko framework is then represented by a valid SMILES string. Finally, the contexts for each molecular pair were represented using standard 1024-bit Daylight fingerprints. For each end point and transformation, the contexts were clustered using the maximum dissimilarity method in Pipeline Pilot,[27] with a minimum Tanimoto similarity of 0.70, thus dividing the contexts into clusters of near-neighbor molecules.

There were two local descriptors: localized RG nodes and atom environments (AEs). Localized RGs are analogous to the global RGs described above, except that, here, only the RG node immediately adjacent to the attachment point is used to characterize the context.[25] Atom environments (AEs) describe a circular neighborhood centered on the attachment point: the descriptor commences with that atom (as encoded by its SYBYL atom type) and then progressively adds the atom types for each layer of atoms up to three bonds away from the attachment point.[28]

Analysis of ΔP Values. After generation of the MMPs, the pairwise property differences were calculated and binned into three bins, indicating the effect of the molecular transformation on the property under examination: a decrease ($\Delta P \leq 0.3$), an increase ($\Delta P \geq 0.3$), and zero (neutral) effect ($-0.3 < \Delta P < 0.3$). The threshold of 0.3 log units was chosen as corresponding to the experimental error of a typical in vitro assay. The labeling was reversed for hERG, where lowering of the inhibition is therapeutically favorable.

For each transformation found in the data set, the distribution of the positive, negative, and zero effects on the property under consideration were summarized by a pie chart or bar chart using “traffic light” coloring (green for favorable, red for unfavorable, amber for neutral).[29] This pie chart or bar chart represents the global probability distribution of the ΔP value after the particular transformation is applied. The context descriptors listed above enable the global distribution for a particular transformation to be analyzed in terms of the many different structural contexts in which that transformation can occur, for example, all occurrences of the transformation H → OMe taking place on a polar aromatic ring (which is one of the localized RG node types). For each specific context, the distribution of positive, neutral, and negative effects on the chosen property was calculated as for the global distribution. Each local distribution was then compared with the corresponding global one, and the statistical significance of the difference (if any) was assessed under the null hypothesis that the local and global distributions were the same. The multinomial test was used to assess the significance of the difference, with the χ^2 test being used to reduce the computation for distributions containing ≥64 data points. A low p value in the test ($p \leq 0.01$) means that there is a significant difference between the global and local distributions, that is, that the transformation is context-sensitive.

RESULTS

The fragment indexing algorithm was applied to the three data sets with the results reported in Tables 2 and 3. Taking hERG as an example in Table 2, the all-pairs comparison of the 76 266 structures yielded a total of 1 431 107 distinct MMPs and 1 035 181 distinct transformations. It has to be noted that the transformations found here do not exactly correspond to the modifications deliberately designed and made by the medicinal chemists during lead optimization. Indeed, the study does not take into account the specifics of each lead optimization project, its scope and targets, chronological account of compound progression, and so on. As would be expected, the distribution of transformation occurrences is extremely skewed, with just 33 of them occurring more than 300 times; Hussain and Rea analyzed transformation frequencies in an MMP analysis of the NIH Molecular Libraries Small Molecule Repository (MLSMR) data and noted that the frequencies follow a Zipfian, power-law distribution.[20] The 33 frequent transformations contained 37 746 distinct transforms, and the number of occurrences for each such frequent transformation ensured that sufficient data were available to derive robust statistics. Turning now to Table 3, part a details the 30 most frequent transformations. It is this transformation subset of the hERG data (and the corresponding subsets of the solubility and lipophilicity data in parts b and c, respectively, of Table 3) that forms the basis for the analyses in the remainder of the article.

Table 2. Application of the Fragment Indexing Algorithm to the Three Data Sets[a]

	hERG	Solubility	Lipophilicity
molecules	76 266	94 053	180 440
distinct MMPs	1 431 107	1 375 382	4 441 033
distinct transforms	1 035 181	927 903	3 169 989
frequent transforms	33	58	175
distinct MMPs in frequent transforms	37 746	56 326	171 040

[a] Frequent transforms are those with more than 300 occurrences within a data set.

DISCUSSION

Analysis of the hERG Data Set. The bar chart in Figure 2 shows the green/amber/red distribution for the 15 most frequent hERG transformations. Three conclusions can be drawn from this figure. First, with the sole exception of the phenyl addition (H → Ph), all of the transformations are chemically very simple, involving the change of just one or two heavy atoms. Given the ubiquitous nature of the transformations, it is hardly surprising that there is a large degree of overlap between these GSK data and the side-chain replacements noted as occurring most frequently in the AstraZeneca study by Haubertin and Bruneau.[12] Second, and more interestingly, amber, denoting a (near-) zero effect, predominates across the bar chart and comprises the largest single portion of the distribution for 9 of these 15 transformations. Thus, these frequent transformations, which happen to be very simple and small, have only a limited effect on hERG inhibition in the majority of cases. However, this is hardly unexpected: the simplicity of the transformations means that the two molecules that are being compared when an MMP is detected will have similar overall structures, and the similar property principle would hence suggest that they exhibit similar property values (i.e., that ΔP is small). Third, those trends that are evident in Figure 2 fit well with existing strategies for modifying hERG inhibition.[22] For example, the most favorable transformation is the addition of a hydroxyl group (H → OH). This results in a significant reduction in hERG inhibition in approximately 45% of the 775 cases, owing to the increased polarity brought about by the hydroxyl addition. The least favorable transformation is

Table 3. Thirty Most Frequent Transformations in Order of Decreasing Number of Examples for (a) hERG Inhibition, (b) Solubility, and (c) Lipophilicity, along with the Percentages of Occurrences Where Each Effect Was Favorable, Unfavorable, or Neutral

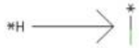
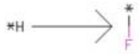
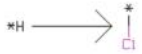
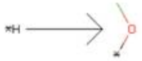
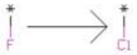
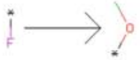
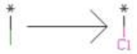
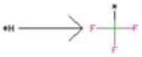
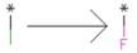
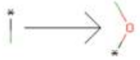
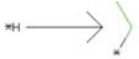
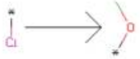
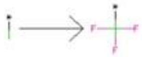
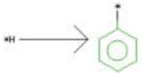
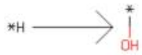
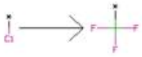
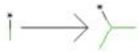
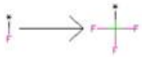
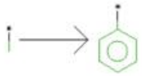
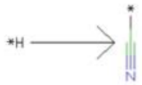
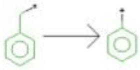
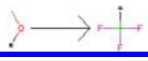
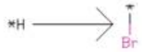
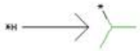
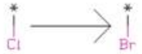
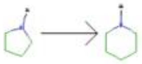
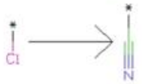
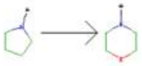
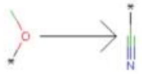
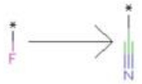
No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
		(a)			
1		8326	18.00	26.59	55.40
2		4243	17.56	23.76	58.68
3		2849	12.71	41.31	45.98
4		2484	25.32	25.16	49.52
5		1441	11.45	34.21	54.34
6		1120	25.54	20.80	53.66
7		1106	12.57	34.45	52.98
8		1094	17.18	42.78	40.04
9		1071	19.89	22.41	57.70
10		1017	29.40	18.19	52.41
11		986	11.26	47.87	40.87
12		927	44.01	14.02	41.96
13		855	18.01	42.22	39.77
14		852	10.33	62.68	27.00
15		775	45.29	10.97	43.74
16		762	21.39	16.67	61.94
17		730	11.78	34.79	53.42
18		625	15.68	41.28	43.04
19		624	12.98	55.29	31.73
20		621	19.65	36.55	43.80
21		616	35.71	16.72	47.56
22		544	18.01	43.57	38.42

Table 3. Continued

(a)

No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
23		510	13.53	48.04	38.43
24		488	11.07	48.57	40.37
25		395	12.41	23.80	63.80
26		372	9.41	26.61	63.98
27		353	36.83	15.30	47.88
28		351	74.64	2.85	22.51
29		344	22.67	30.23	47.09
30		339	20.06	28.91	51.03

(b)

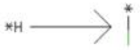
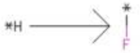
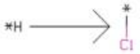
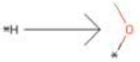
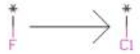
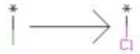
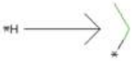
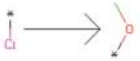
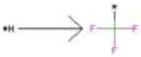
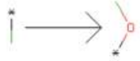
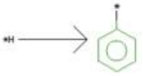
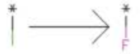
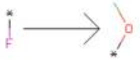
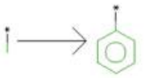
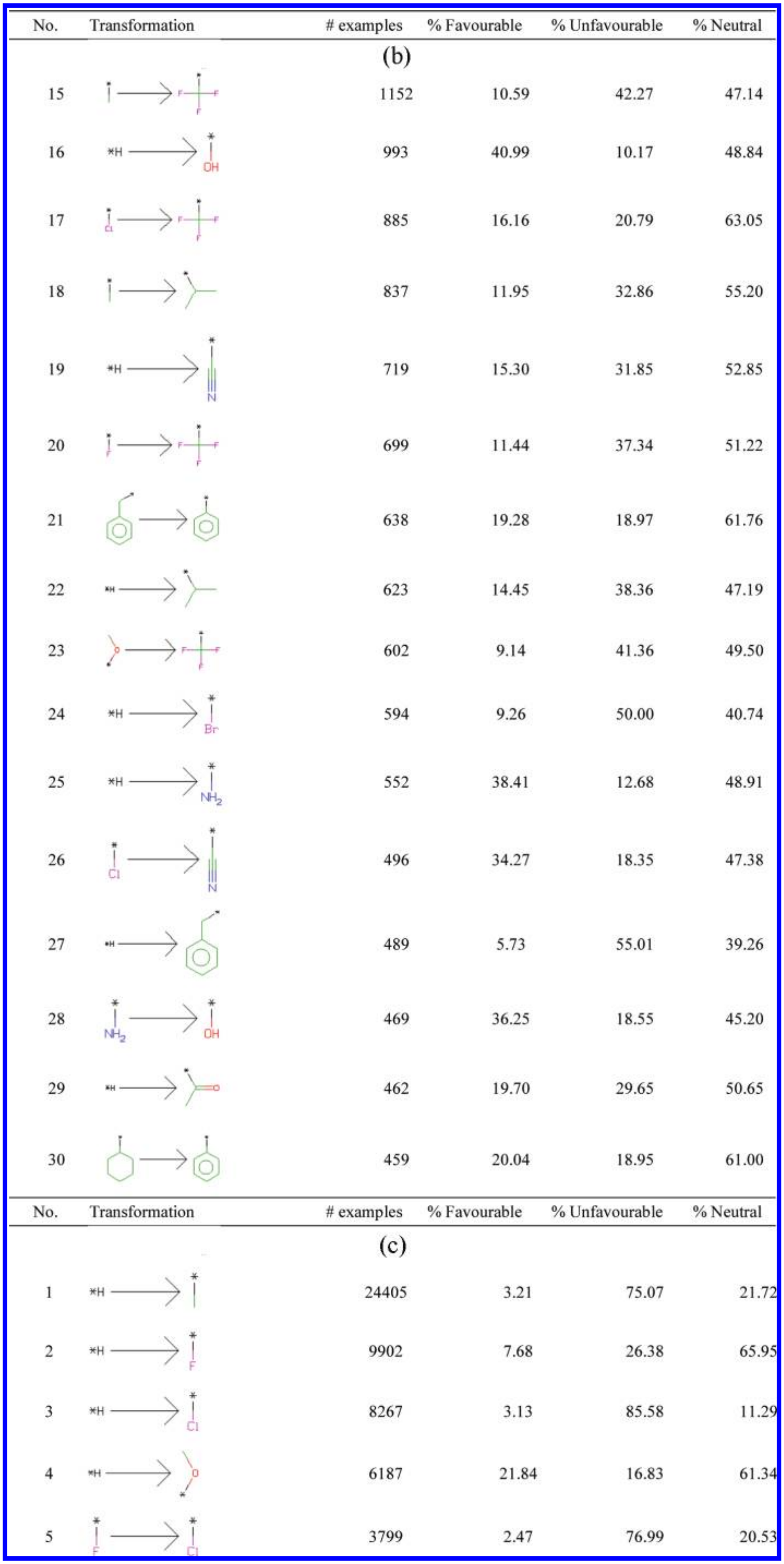
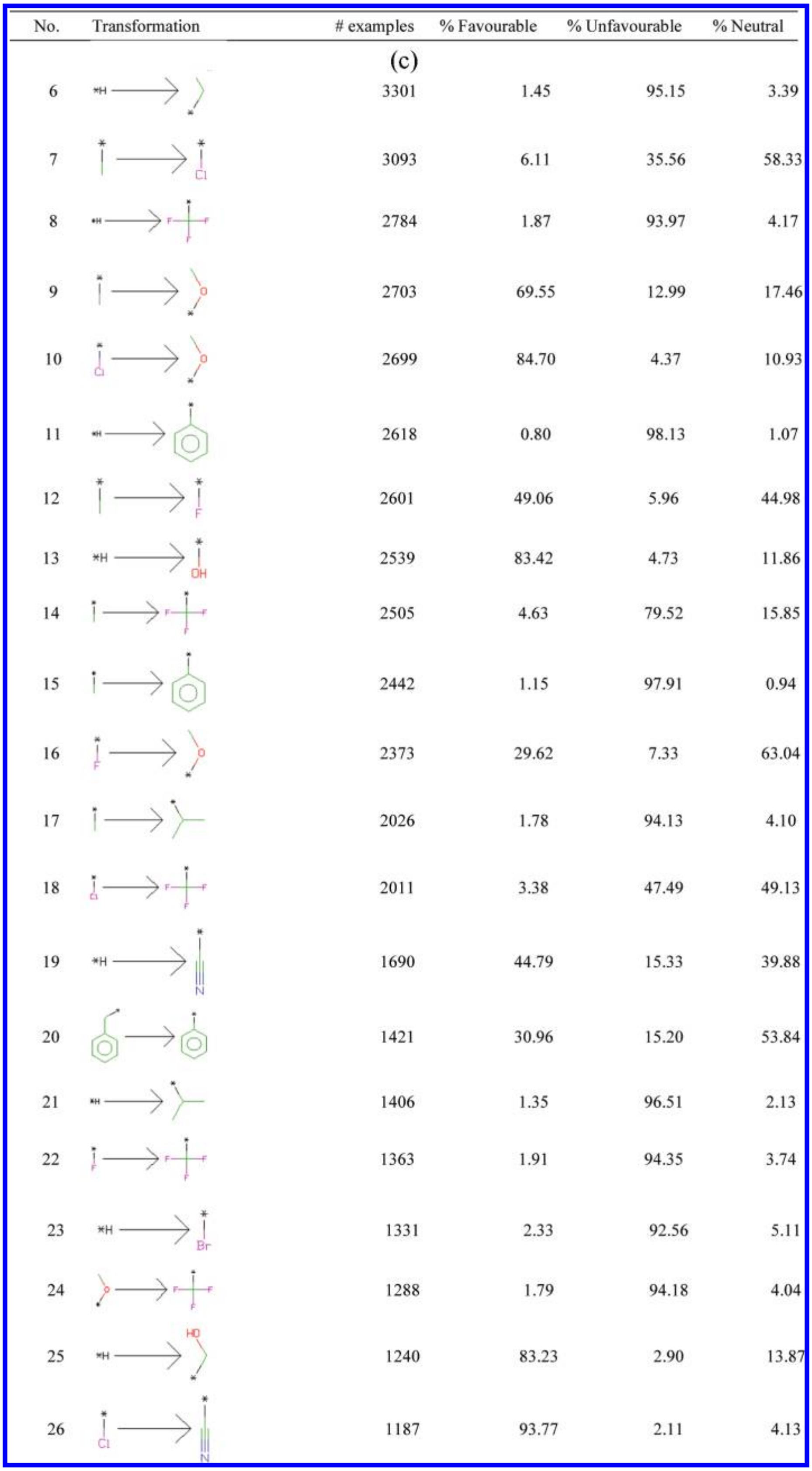
No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
1		10695	15.18	24.95	59.87
2		4273	16.50	20.17	63.33
3		4037	10.87	38.25	50.88
4		3104	18.78	19.88	61.34
5		1676	11.34	35.74	52.92
6		1566	12.45	29.69	57.85
7		1363	11.96	39.77	48.28
8		1359	38.04	9.79	52.17
9		1351	10.58	46.93	42.49
10		1323	25.77	13.08	61.15
11		1288	6.13	61.57	32.30
12		1216	21.30	17.68	61.02
13		1163	22.27	15.39	62.34
14		1158	6.22	57.08	36.70

Table 3. Continued



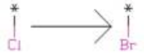
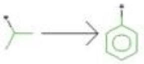
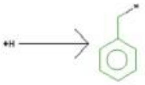
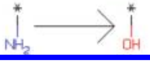
No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
		(b)			
15		1152	10.59	42.27	47.14
16		993	40.99	10.17	48.84
17		885	16.16	20.79	63.05
18		837	11.95	32.86	55.20
19		719	15.30	31.85	52.85
20		699	11.44	37.34	51.22
21		638	19.28	18.97	61.76
22		623	14.45	38.36	47.19
23		602	9.14	41.36	49.50
24		594	9.26	50.00	40.74
25		552	38.41	12.68	48.91
26		496	34.27	18.35	47.38
27		489	5.73	55.01	39.26
28		469	36.25	18.55	45.20
29		462	19.70	29.65	50.65
30		459	20.04	18.95	61.00
No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
		(c)			
1		24405	3.21	75.07	21.72
2		9902	7.68	26.38	65.95
3		8267	3.13	85.58	11.29
4		6187	21.84	16.83	61.34
5		3799	2.47	76.99	20.53

Table 3. Continued



No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
		(c)			
6		3301	1.45	95.15	3.39
7		3093	6.11	35.56	58.33
8		2784	1.87	93.97	4.17
9		2703	69.55	12.99	17.46
10		2699	84.70	4.37	10.93
11		2618	0.80	98.13	1.07
12		2601	49.06	5.96	44.98
13		2539	83.42	4.73	11.86
14		2505	4.63	79.52	15.85
15		2442	1.15	97.91	0.94
16		2373	29.62	7.33	63.04
17		2026	1.78	94.13	4.10
18		2011	3.38	47.49	49.13
19		1690	44.79	15.33	39.88
20		1421	30.96	15.20	53.84
21		1406	1.35	96.51	2.13
22		1363	1.91	94.35	3.74
23		1331	2.33	92.56	5.11
24		1288	1.79	94.18	4.04
25		1240	83.23	2.90	13.87
26		1187	93.77	2.11	4.13

Table 3. Continued

No.	Transformation	# examples	% Favourable	% Unfavourable	% Neutral
		(c)			
27		1134	4.14	9.88	85.98
28		1027	18.31	40.90	40.80
29		1018	0.98	98.13	0.88
30		1006	62.43	26.94	10.64

the addition of a phenyl ring (H → Ph). This transformation results in a significant increase in hERG inhibition in approximately 65% of the 852 cases, owing to the increased lipophilicity brought about by the addition of the featureless aromatic ring.

Figure 2 describes the global, context-independent situation. However, the use of the context descriptors outlined in Methods demonstrates multiple statistically significant trends (some positive and some negative) that are not apparent when the entire data set is considered. In all, we identified a large number of cases that gave a significantly different ($p \leq 0.01$) distribution from the global distribution, as listed in Table 4. Here, we discuss three examples to illustrate the enhanced predictive power of the MMP approach when contextual information is available.

In the first example (Figure 3), the global distribution of the ΔhERG values for the H → OCH_3 transformation (for which there are 2484 examples available for analysis) indicates that the chance of reducing hERG inhibition is approximately 25% and almost exactly the same as the chance of increasing it. If, however, account is taken of the RG node immediately adjacent to the attachment point, then a more complex pattern of behavior becomes evident. If one considers the 161 cases in which the transformation takes place adjacent to an aliphatic linker, then there is a much-increased chance of reducing hERG, whereas the opposite effect is clearly evident in the 108 cases where the transformation takes place next to an H-bond-accepting aromatic ring (such as pyridine or quinoline). However, if the context is a featureless aromatic ring (the vast majority of the cases), then the distribution is very similar to the global distribution, namely, hERG inhibition is likely to remain relatively unaffected when the methoxy group is added. In the second example (Figure 4), the global ΔP distribution suggests that the transformation CH_3 → F is as likely to result in an increase in hERG as it is to result in a decrease. However, an increase in hERG inhibition is strongly indicated in the presence of the particular AE context shown in the figure. Finally, Figure 5 demonstrates that hERG inhibition is also likely to be increased when the listed Murcko framework provides the local context for the cyclohexyl → phenyl transformation.

It is possible to rationalize the observed behavior in all three cases. In Figure 3, the presence of an aliphatic chain enhances the polarity and the hydrogen-bonding capabilities of the oxygen atom of the added methoxy group, with a consequent increase in the chances of reducing hERG. Conversely, the addition of the same methoxy group to an aromatic ring reduces the hydrogen-bonding capabilities of the oxygen atom, with the extra hydrophobic methyl increasing lipophilicity and thus hERG inhibition. In Figure 4, the vast majority of the 117 contexts belong to a single

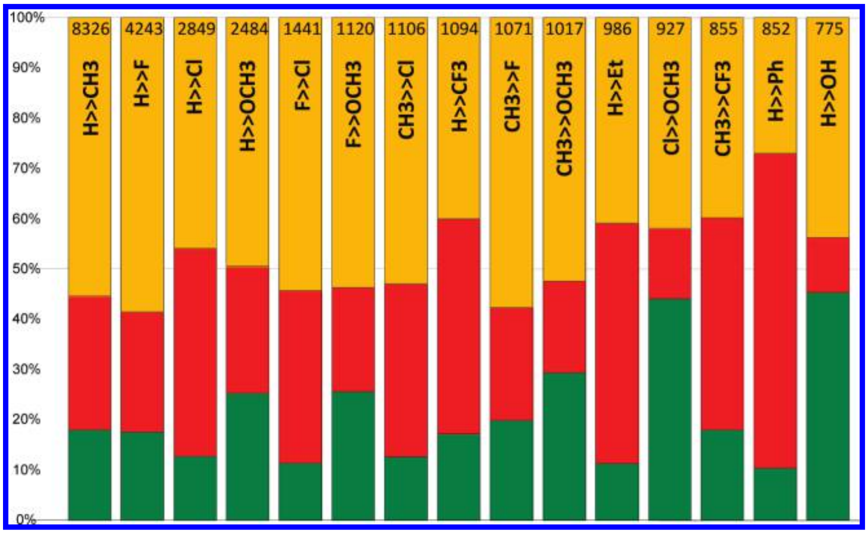


Figure 2. Bar chart for the 15 most frequent transformations in the hERG data set. The number at the top of each bar is the total frequency of occurrence for the listed transformation. The colors reflect the therapeutic effect of each transformation with red, amber, and green denoting unfavorable (increase), zero, and favorable (decrease) changes, respectively, in hERG inhibition.

Table 4. Numbers of Statistically Significant ($p \leq 0.01$) Contexts Identified Using the Five Different Types of Descriptors

descriptor	hERG	solubility	lipophilicity
Daylight fingerprints	187	367	1220
Murcko frameworks	159	243	1242
reduced graphs	165	320	1329
atom environments	229	274	1719
localized reduced graph nodes	32	91	439

chemotype. The substitution of a methyl with a smaller fluorine in the ortho position is likely to induce a slight conformational change in the ligand and renders the bioactive conformation more accessible, thus increasing the chances of hERG inhibition. Finally, in Figure 5, we ascribe the increase in hERG inhibition to three factors: the substitution of the cyclohexyl ring by its aromatic equivalent leads to a slight increase in lipophilicity; the extra aromatic ring increases the planarity of the molecule and, thus, its hydrophobicity; and the phenyl group reduces the polarity of the amidic carbonyl adjacent to it and, hence, its ability to accept hydrogen bonds.

As these three examples demonstrate, we have found that it is normally possible to rationalize any differences that are observed between the global (i.e., context-independent) and local (i.e., context-sensitive) distributions; however, many of the differences are ones that might not have been immediately obvious to the medicinal chemist engaged in an optimization project. The inclusion of contextual information hence provides an additional source of information to assist the chemist in deciding which analogues should be synthesized next in the search for a viable candidate. A similar conclusion can be drawn from our studies of the solubility and lipophilicity data sets, as we now demonstrate.

Analysis of the Solubility and Lipophilicity Data Sets. The bar chart in Figure 6 shows the 16 most frequent solubility transformations. The trends displayed here are similar to those in Figure 2, in that most of the frequent transformations involve just one or two heavy atoms; indeed, all but one of the transformations are present in both bar charts, and the zero effect (amber) is again the most common of the three types of effects.

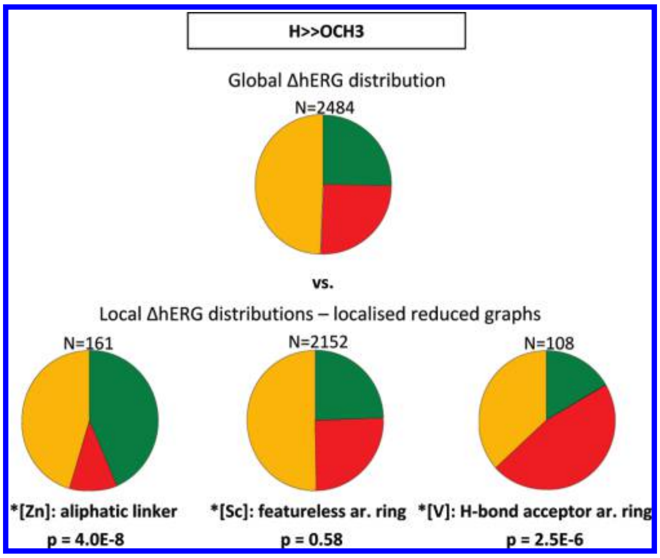


Figure 3. Global and local ΔP distributions for the H → OCH_3 transformation in the hERG data set. Colors as in Figure 2. Different trends are observed, depending on whether the reduced graph node of the attachment point is an aliphatic linker [Zn], a hydrophobic aromatic ring [Sc], or a polar aromatic ring [V]. P values signify the statistical significance of this observation. The number of examples for each case is shown above the respective pie chart.

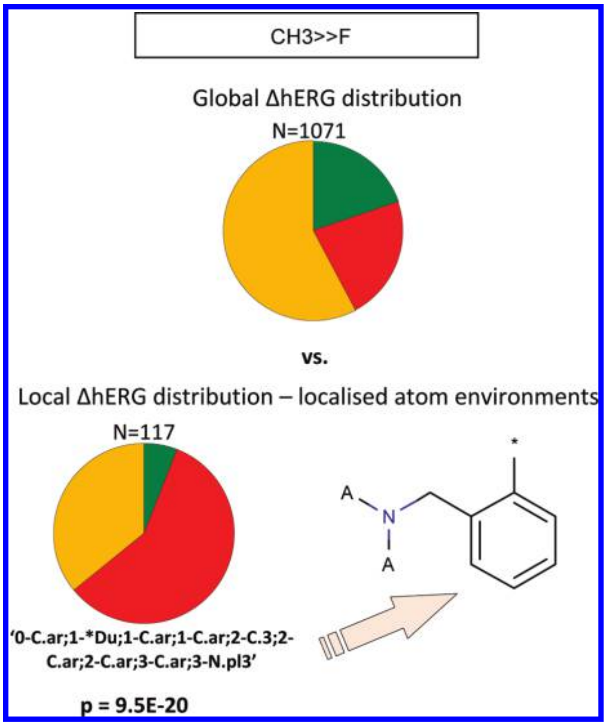


Figure 4. Global and local ΔP distributions for the CH_3 → F transformation in the hERG data set. Colors as in Figure 2. The subset of 117 contexts has the same environment around the attachment point, as identified by the localized AE descriptor (where A represents any non-hydrogen atom).

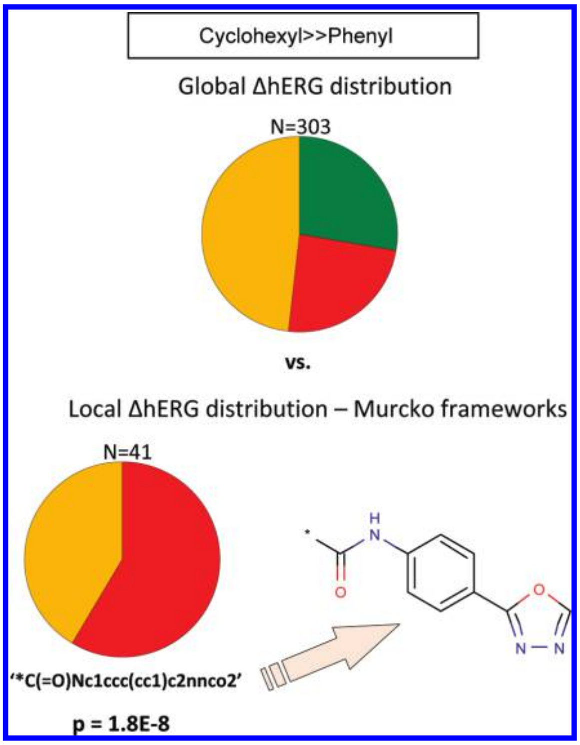


Figure 5. Global and local ΔP distributions for the cyclohexyl → phenyl transformation in the hERG data set. Colors as in Figure 2. The subset of 41 contexts has the same environment around the attachment point, as identified by the Murcko framework on the right.

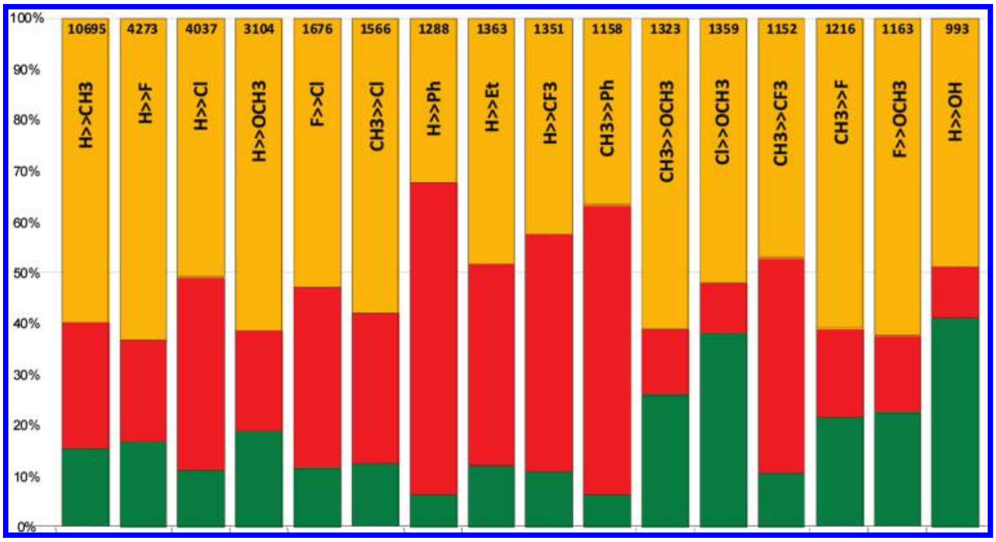


Figure 6. Bar chart for the 16 most frequent transformations in the solubility data set. The number on the top of each bar is the frequency of occurrence for the listed transformation. The colors reflect the effect of each transformation with red, amber, and green denoting unfavorable (decrease), zero, and favorable (increase) changes, respectively, in solubility.

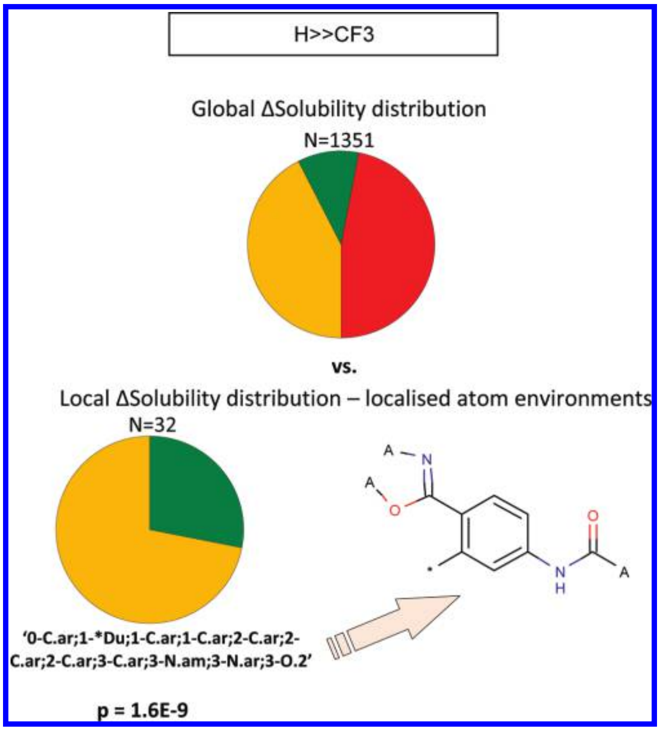


Figure 7. Global and local ΔP distributions for the H → CF_3 transformation in the solubility data set. Colors as in Figure 6. The subset of 32 contexts has the same environment around the attachment point, as identified by the localized AE descriptor (where A represents any non-hydrogen atom).

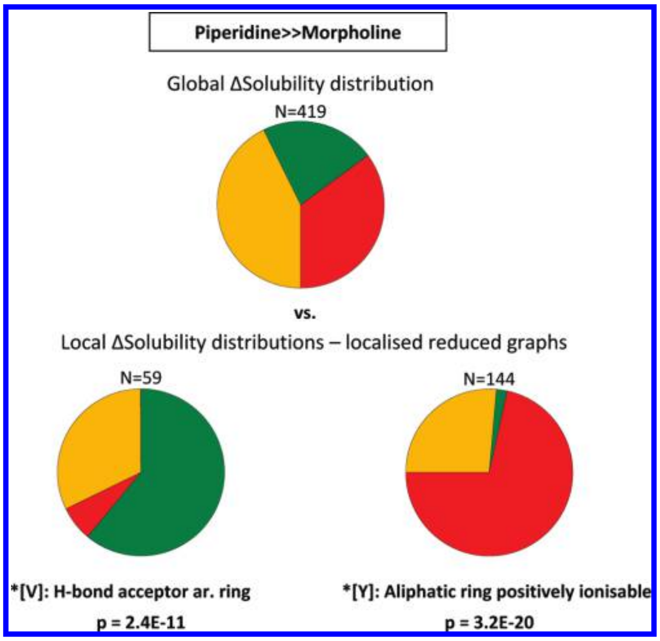


Figure 8. Global and local ΔP distributions for the piperidine → morpholine transformation in the solubility data set. Colors as in Figure 6. Different trends are observed, depending on whether the reduced graph node of the attachment point is a polar aromatic ring [V] or a positively ionizable aliphatic ring [Y].

At the global level, the transformation H → OH is the best way of increasing solubility, as a result of the increased polarity caused by the presence of the hydroxyl group; conversely, a decrease in solubility is most likely as a result of adding a hydrophobic ring (H → Ph). These observations are unremarkable; less obvious are the trends evident in Figures 7−9. The global ΔS distribution in the first example (H → CF_3) is certainly not unexpected, because the addition of the lipophilic $-CF_3$ group has a clear negative effect on solubility. However, very different behavior is observed when this transformation takes place with the context defined by the AE descriptor shown in Figure 7. We ascribe this to two factors: (a) the addition of a $-CF_3$ group in the ortho position breaks the coplanarity and hence reduces the crystal packing, thus ultimately favoring solubility, and (b) the addition of a $-CF_3$ group enhances the H-bond acidity of the amidic nitrogen, rendering it more polar. The nature of the RG node next to the attachment point has a marked effect on the piperidine → morpholine transformation, as demonstrated in Figure 8: If the local context is an H-bond accepting aromatic ring, then solubility is likely to increase when this transformation is carried out, whereas a decrease is to be expected when an aliphatic positively ionizable ring (e.g., pyrrolidine) provides the context. The final example (Cl → OCH_3) in Figure 9 involves clustering of the contexts using Daylight fingerprints. In this figure, the points illustrate the contexts belonging to a specific cluster. Each context is derived from a matched pair, which is, in turn, linked to a

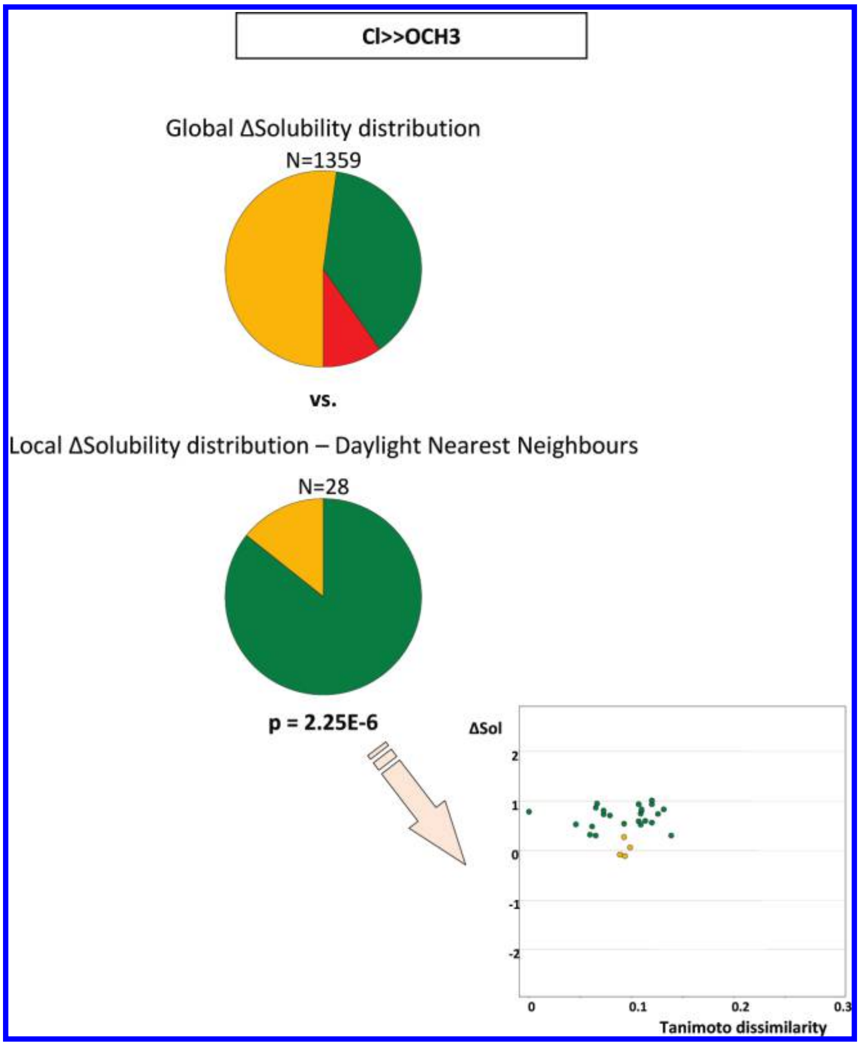


Figure 9. Global and local ΔP distributions for the Cl → OCH_3 transformation. Colors as in Figure 6. The 28 contexts comprising the second pie chart are Daylight nearest neighbors, and they exhibit a solubility behavior that is very different from the global behavior. The scatter plot on the right is the dissimilarity−ΔP plot for the particular cluster, using the cluster center as the reference.

transformation and the associated ΔP value. The x axis of the figure is the Tanimoto dissimilarity to the cluster center, which is used as the reference, and the y axis is the resulting difference in solubility (ΔP) associated with each data point. The global ΔP distribution is the expected one, given the substitution of a lipophilic fragment for a less lipophilic one. In the subset of 28 contexts, however, the impact of the substitution is much more evident, leading to an increase in solubility in 85% of the 28 cases. The vast majority of the contexts in that cluster have a basic center near the attachment point. Therefore, the increase in solubility should be due to the increase in basicity, caused by the substitution of an electron-withdrawing group (Cl) by an electron-donating one (OCH_3).

The bar chart in Figure 10 shows the 21 most frequent lipophilicity transformations. Optimization generally seeks to achieve a reduction in log D (because this can counter promiscuity, toxicity, hERG inhibition, and insolubility),[30] and increased log D is hence here colored in red (i.e., a negative effect). The majority of the most frequent transformations are again small and simple, but a comparison with the corresponding Figures 2 and 6 demonstrates clearly the much greater scope for transformations with an unfavorable outcome on the target property. For example, the addition of an ethyl, phenyl, or isopropyl group will lead to an increase in log D value in more than 95% of the cases. Although these examples are unsurprising, the prevalence of negative effects here and in Table 3c makes very clear the challenges facing the medicinal chemist who needs to reduce log D. That said, there are obvious examples in the figure that can be expected to have a positive outcome, such as the substitution of a lipophilic chloride by a polar methoxy group or the addition of a polar hydroxyl group. Two less obvious distributions are considered further below.

The global distribution in the upper part of Figure 11 shows that the H → nitrile transformation results in a decrease in log D value in almost half of the 1690 cases, which is not unexpected given the polarity of the nitrile group. Very different behavior is demonstrated in the lower part of the figure. On the left, the nitrile is added next to an H-bond acceptor aromatic ring, such as a pyridine or a pyrimidine. The strong electron-withdrawing property of the nitrile group reduces the hydrogen-bonding potential (log K_β) of the heteroaryl nitrogen, thus increasing log D. Conversely, the addition of the nitrile group on an aliphatic linker has more expected effects, as depicted by the pie chart on the right. Figure 12 describes the isopropyl → phenyl transformation, with the global distribution suggesting a tendency toward increased log D due to the introduction of the bulkier phenyl ring. The use of the AE descriptor shows two

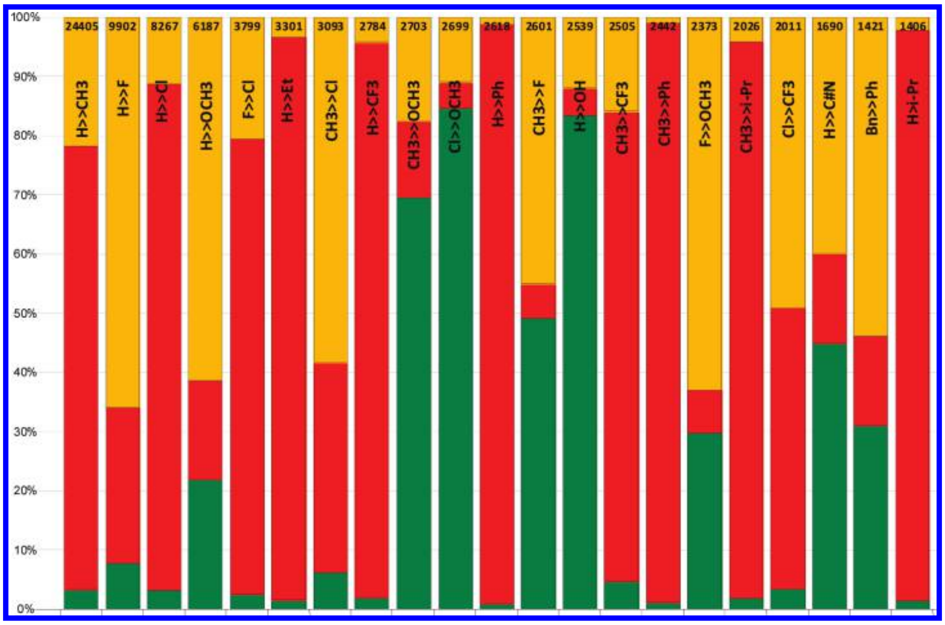


Figure 10. Bar chart for the 21 most frequent transformations of the lipophilicity data set. The number at the top of each bar is the total frequency of occurrence for the listed transformation. The colors reflect the effect of each transformation with red, amber, and green denoting unfavorable (increase), zero, and favorable (decrease) changes, respectively, in lipophilicity.

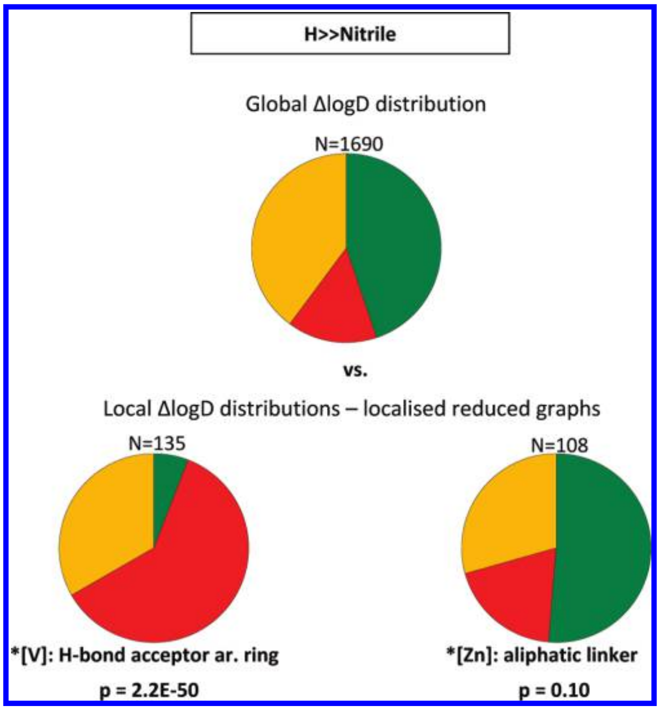


Figure 11. Global vs local Δ*P* distributions for the H → C≡N transformation. Colors as in Figure 10. The subsets of 135 and 108 contexts have the same reduced graph node, an aromatic ring H-bond acceptor and an aliphatic linker, respectively.

dramatically different types of behavior in the lower part of the figure. On the left, the transformation takes place next to a methylene-oxy-aryl moiety and results in a decrease in lipophilicity in ca. 70% of the 152 examples with this particular context. It is not trivial to rationalize this observation, but it is demonstrably a meaningful one, with the 152 examples covering more than 25 distinct Murcko frameworks. This is also the case on the left, where the addition of the phenyl ring near a heteroaryl moiety has no substantial effect on log *D* for ca. 90% of the 41 cases.

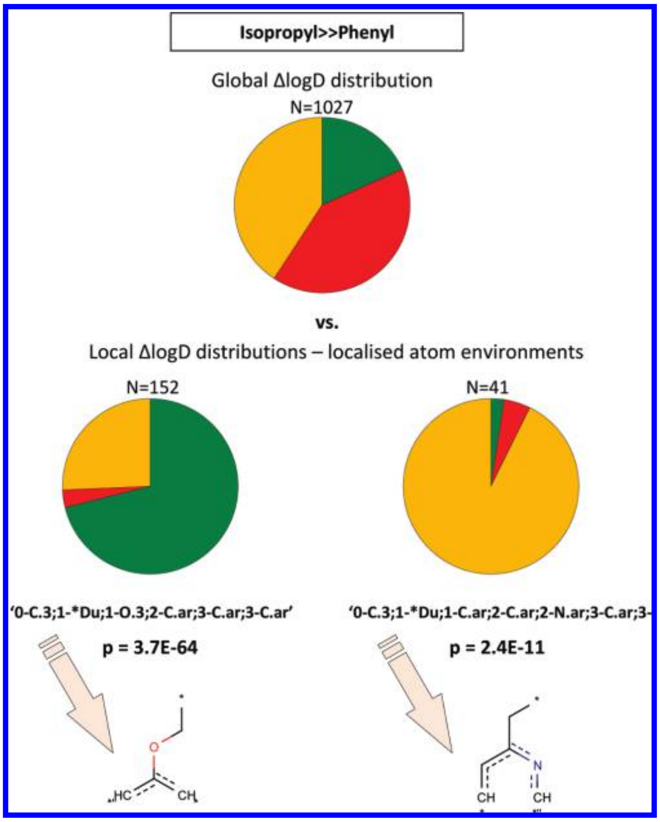


Figure 12. Global vs local Δ*P* distributions for the isopropyl → phenyl transformation. Colors as in Figure 10. The subsets of 152 and 41 contexts have the same environments around the attachment point, as identified by the localized AE descriptor.

General Considerations. In concluding this section, we make two general points about the natures of the contexts and about the descriptors we have chosen to probe the contexts.

First, it must be emphasized that the method detailed here is statistical in nature: a low *p* value simply indicates a context where the local distributions are significantly different from the global one and does not necessarily mean that this

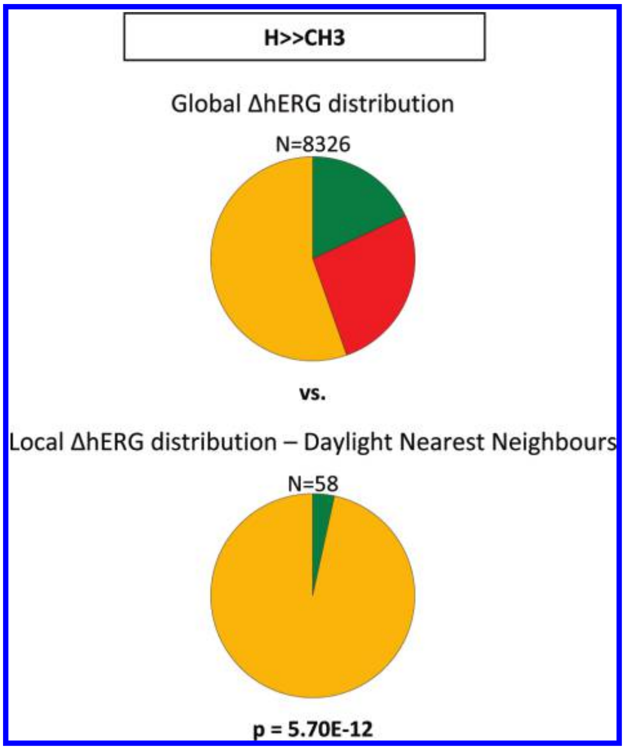


Figure 13. Global and local distributions for methyl addition in the hERG data set. Colors as in Figure 2. The 58 contexts comprising the second pie chart are Daylight nearest neighbors, and they clearly exhibit a behavior that is very different from the global one.

discrepancy is of potential value in medicinal chemistry terms. For example, Figure 13 shows that the addition of a methyl group (H → CH_3) has a nonpositive hERG effect for ca. 80% of the 8326 examples of this transformation. However, there is a large cluster of molecules (all belonging to the same chemotype), where this is not the case on methyl addition, as indicated by the Daylight nearest-neighbor analysis. Whether this is useful, or even interesting, will depend on the particular problem that the medicinal chemist needs to address. That said, the previous examples clearly demonstrate that the method provides an effective, knowledge-based idea generator that could suggest optimization strategies that would not otherwise have been considered.

The descriptors used here provide a hierarchical way to look at a context, starting from the whole structure and focusing progressively on the localized environment where the transformation takes place. Thus, the use of Daylight fingerprints or Murcko frameworks will lead to identification of different chemotypes or series within the data, with the latter providing a less specific, and hence more abstract, description of a chemotype than do the Daylight fingerprints. The localized descriptors, such as the adjacent reduced graph node or the atom environment, do not consider chemotypes at all but provide a more focused encoding of the substructural neighborhood. Specifically, the atom environment descriptors provide more focused representations than does the reduced graph node, which does not encode precise information on connectivity and atom types. As always, there is a tradeoff between specificity and generalizability: the more specific the description of the context, the more difficult it becomes to extrapolate. Hence, the choice of context descriptor will be conditioned by the requirements at a specific stage in a lead optimization project.

CONCLUSIONS

The past few years have seen considerable interest in the use of MMPs to assist the medicinal chemist in the lead optimization stage of drug discovery, providing guidance as to the possible property changes resulting from the use of a particular type of synthetic transformation. This information can then be used to decide which analogues should be synthesized next in the optimization. Conventional approaches to the analysis of MMPs often take no account of the structural context in which a transformation takes place. In this article, we have demonstrated that such approaches can provide only partial, and in some cases highly misleading, guidance as to the changes in property that might result. The use of contextual information, as encoded here using several different types of substructural descriptors, can provide a much more nuanced appraisal of the effect of a particular transformation for a particular research program.

Having demonstrated the power of contextual information, our current work focuses on the use of additional types of substructural descriptors and the development of a decision support tool to provide a routine source of assistance to the medicinal chemist. This work will be reported shortly.

ACKNOWLEDGMENT

We thank the Engineering and Physical Sciences Research Council and GlaxoSmithKline for funding. G.P. also thanks the Alexander S. Onassis Public Benefit Foundation in Greece for additional financial support.

REFERENCES AND NOTES

(1) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics. 2nd ed.*; Kluwer: Dordrecht, The Netherlands, 2007.
(2) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(3) Holliday, J. D.; Jelfs, S. P.; Willett, P. Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406–411.
(4) Wagener, M.; Lommerse, J. P. M. The quest for bioisosteric replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.
(5) Krier, M.; Hutter, M. C. Bioisosteric similarity of molecules based on structural alignment and observed chemical replacements in drugs. *J. Chem. Inf. Model.* **2009**, *49*, 1280–1297.
(6) Birchall, K.; Gillet, V. J.; Willett, P.; Ducrot, P.; Luttmann, C. Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1330–1346.
(7) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178–180.
(8) Free, S. M.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395–399.
(9) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular-Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(10) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. Chem-Bioinformatics: Comparative QSAR at the Interface between Chemistry and Biology. *Chem. Rev.* **2002**, *102*, 783–812.
(11) Lewis, R. A. A general method for exploiting QSAR models in lead optimization. *J. Med. Chem.* **2005**, *48*, 1638–1648.
(12) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(13) Sheridan, R. P.; Hunt, P.; Culbertson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
(14) 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.
(15) 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.

(16) 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.

(17) 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.

(18) Southall, N. T.; Ajay. Kinase patent space visualization using chemical replacements. *J. Med. Chem.* **2006**, *49*, 2103–2109.

(19) 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.

(20) 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.

(21) 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.

(22) Jamieson, C.; Moir, E. M.; Rankovic, Z.; Wishart, G. Medicinal chemistry of hERG optimizations: Highlights and hang-ups. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(23) Kerns, E. H.; Di, L. *Drug-Like Properties: Concepts, Structure Design and Methods: from ADME to Toxicity Optimization*; Academic Press: San Diego, CA, 2008.

(24) Bhattachar, S. N.; Wesley, J. A.; Seadeek, C. Evaluation of the chemiluminescent nitrogen detector for solubility determinations to support drug discovery. *J. Pharm. Biomed. Anal.* **2006**, *41*, 152–157.

(25) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145–2156.

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

(27) *Accelrys Pipeline Pilot*, version 7.0; Accelrys Inc.: San Diego, CA, 2010.

(28) 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.

(29) Lobell, M.; Hendrix, M.; Hinzen, B.; Keldenich, J.; Meier, H.; Schmeck, C.; Schohe-Loop, R.; Wunberg, T.; Hillisch, A. In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* **2006**, *1*, 1229–1236.

(30) Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discovery* **2010**, *5*, 235–248.

CI100258P