Bioorganic & Medicinal Chemistry 19 (2011) 3739–3749

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Extraction of tacit knowledge from large ADME data sets via pairwise analysis

Christopher E. Keefer [a,*], George Chang [a], Gregory W. Kauffman [b]

[a] Computational ADME Group, Department of Pharmacokinetics, Dynamics, and Drug Metabolism, Pfizer Inc., Groton, CT 06340, USA
[b] Worldwide Medicinal Chemistry, Neuroscience Research Unit, Pfizer Inc., Groton, CT 06340, USA

ARTICLE INFO

Article history:
Received 10 March 2011
Revised 28 April 2011
Accepted 3 May 2011
Available online 6 May 2011

Keywords:
Pairwise
ADME
Activity cliff
Switch
HLM
Permeability
MDCK
P-gp
MDR1
Lipophilicity
Log *D*
SAR

ABSTRACT

Pharmaceutical companies routinely collect data across multiple projects for common ADME endpoints. Although at the time of collection the data is intended for use in decision making within a specific project, knowledge can be gained by data mining the entire cross-project data set for patterns of structure–activity relationships (SAR) that may be applied to any project. One such data mining method is pairwise analysis. This method has the advantage of being able to identify small structural changes that lead to significant changes in activity. In this paper, we describe the process for full pairwise analysis of our high-throughput ADME assays routinely used for compound discovery efforts at Pfizer (microsomal clearance, passive membrane permeability, P-gp efflux, and lipophilicity). We also describe multiple strategies for the application of these transforms in a prospective manner during compound design. Finally, a detailed analysis of the activity patterns in pairs of compounds that share the same molecular transformation reveals multiple types of transforms from an SAR perspective. These include bioisosteres, additives, multiplicatives, and a type we call switches as they act to either turn on or turn off an activity.

1. Introduction

While innovation is essential in drug discovery, the capture and exploitation of tacit or soft knowledge that led to previously successful drug candidates remain critical endeavors for the pharmaceutical industry in its quest to meet the increasing demands to produce new drug candidates. The capture of this tacit knowledge, whether internal or external, is complex as it must be codified for conversion to explicit or organizational knowledge.[1] More challenging than compilation is the mining of this knowledge for the right information at the appropriate time. Within the realm of analog design, the current cumulative knowledge of known and yet to be deciphered medicinal chemistry principles is arguably embedded in the properties and activities of prior compounds. Different strategies have been pursued to make use of this data, many with orthogonal focus. Statistical QSAR models, while able to score novel compounds relative to each other, lack the ability to suggest chemical modifications for a desired activity change. By contrast, pairwise analysis has the potential to yield design ideas that result in a desired activity change. A matched molecular pair is a pair of compounds that differ only by a relatively small structural feature change. The structural transformation in a matched molecular pair represents the chemical modification associated with all activity differences between the pair. Hence, mining pairwise transformations and their corresponding activities may provide chemical modification solutions for a particular activity change. Unknown is whether the change corresponding to the transformation is unique to that particular pair or is a more general phenomenon. Multiple examples of the same transform, coupled with statistical analysis, can provide context and confidence to the generality of the change. Consequently, a comprehensive catalog of pairwise transforms coupled with an efficient search algorithm would be a valuable tool for data mining resulting in a codification of (tacit) medicinal chemistry knowledge.

Several approaches for pairwise analysis have been reported in the literature.[2–7] Two critical factors for effective pairwise analysis are an efficient algorithm for generating comprehensive matched molecular pair lists and large activity data sets for generating confidence in the resulting patterns. Early approaches for generating comprehensive matched molecular pair lists were limited by the computational expense of the maximum common subgraph (MCS) algorithm. Recently, an efficient algorithm to identify matched molecular pairs was reported in the literature which overcomes many of the computational liabilities of the traditional approaches.[8,9] The computational efficiency of this algorithm enables comprehensive pairwise analysis of data sets that could only be approximated previously. The confluence of this new algorithm,

* Corresponding author. Tel.: +1 860 686 4842.
E-mail address: christopher.keefer@pfizer.com (C.E. Keefer).

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.05.003

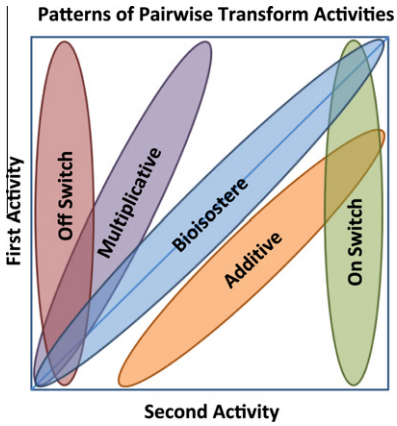


Figure 1. Types of patterns seen in transform activity graphs. For the matched molecular pairs in the transform A → B, the activities of the compounds containing the A fragment are on the Y-axis (First Activity) and the activities of the compounds containing the B fragment are on the X-axis (Second Activity). These patterns can occur above or below the diagonal.

our large internal database of ADME (absorption, distribution, metabolism and excretion) endpoints and the recognition that during the design phase of projects, simultaneous optimization of multiple ADME and primary activity endpoints is required, prompted our development of a streamlined process to codify and continuously mine soft-knowledge transformations from the Pfizer ADME data.

In this paper, we describe the process of cataloging our ADME databases for pairwise transformations and mining the transformation tables for the appropriate knowledge. Multiple approaches for mining the data are possible and three distinct tactics will be highlighted. The first is to present a compound or a specific substructure and identify all existing transformations which may provide the desired activity change thereby yielding possible chemical modification ideas or solutions. The second is to search all activity change knowledge for a particular transformation thereby providing context of whether the chemical change is beneficial or detrimental across multiple ADME endpoints. The third is to mine the pairwise database for known or yet to be deciphered medicinal chemistry principles in ADME space. This is done via analysis of the patterns that emerge in plots of the pairwise activities against each other for a given transform. Figure 1 shows several different types of patterns we have identified in our ADME data sets. These patterns are complex and demonstrate that small molecular transformations can have very different effects that go beyond simple additive activity change. The patterns which reflect bioisosteres and additives are clear and well understood. The other patterns that reflect multiplicative or switch-like phenomena are less intuitive. The importance and implications of these patterns and their use in design will be presented. Ultimately, mining these databases should provide a means to exploit the embedded tacit knowledge and extract the appropriate ADME solution for the problem at hand.

2. Material and methods

2.1. Matched molecular pair identification

We have implemented a modified version of the algorithm published by Hussain and Rea,[8] which we call PairFinder, in C++ using the OEChem Toolkit.[10] The algorithm takes as input a list of SMILES strings with associated activity values and identifies all matched molecular pairs. Options include the number of allowed R-groups in a fragment, along with fragment size limits for terminal fragments and core fragments. Several files can be generated as output. First, a file of individual matched molecular pairs (one per line) can be generated along with the activity change for the pair. In addition, a file of summarized information for pairs that have the same molecular transform can also be generated. All summary transforms are output in the direction of increasing change to reduce the overall size of the generated file. The summary file also includes the number of pairs and the mean, standard deviation, and *p*-value of the activity changes for the transform (using a matched pairs Student's t-test). The percentages of pairs with activities that increase, decrease, or stay the same (less than 20% change) are also generated. In order to identify regioisomeric matched molecular pairs, R-group labels for core fragments are tracked. This requires identification and normalization of cases where there are equivalent R-groups. Equivalent R-groups fall into two categories: those where the R-groups are the same for a given pair and those where the substitution positions of the R-groups on the core are equivalent (a disubstituted phenyl ring, for example). Figure 2 shows an example for each of these cases. Calculation times for PairFinder running on a single CPU are very good and scale roughly linearly with the number of compounds in the data set (~2,000 compounds/s). In order to keep the information gained by PairFinder up-to-date and relevant to current projects, the process of building these pairwise databases has been automated and the databases are updated on a regular basis.

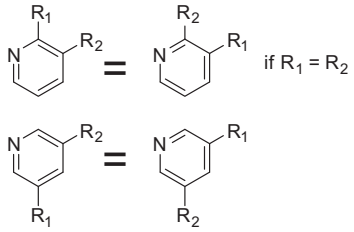


Figure 2. Types of R-group equivalence that need normalization in the PairFinder method.

2.2. Application of transforms for idea generation

A complementary application to PairFinder called PairTransformer has also been implemented. The purpose of this algorithm is to apply transforms generated by PairFinder to user input molecules, thus creating new molecules which can be used for idea generation. The application is written in C++ and the OEChem toolkit.[10] It takes as input a set of structures with or without defined substructures and then fragments these structures/substructures using the same rules as PairFinder. Each of the generated fragments is searched against the summary file produced by PairFinder and matching transforms are applied to generate new ideas (compounds). PairTransformer has optional filters for the transforms it will apply including the magnitude of the mean activity change, minimum number of pairs, and whether the change is increasing, decreasing, both, or neither. This functionality has been integrated with an in-house application called the Pfizer Compound Analysis Tool (PCAT)[11] to facilitate use by project teams.

2.3. Data sets

Four data sets from Pfizer's high-throughput ADME screening efforts[12] were analyzed for this paper: human liver microsomal apparent intrinsic clearance (HLM), passive membrane permeability in a MDCK cell-line (RRCK), P-gp efflux (BA/AB permeability ratio) in the MDR1 cell line (MDR), and lipophilicity using the shake-flask log *D* method (SF Log *D*). The HLM data set contains

226,348 compounds with an effective range of 8–400 μL/min/mg. The RRCK data set contains 102,933 compounds with an effective range of 2–50 $\times 10^{-6}$ cm/s. The MDR efflux data set contains 74,624 compounds with effective range of ~1–100 and is unitless. The SF Log *D* data set contains 29,998 compounds with an effective range of −1.5 to 4.0 and is unitless. Data that were outside the limits of detection were given the limit value. PairFinder was run on all four data sets with the same options: up to three bond breaks for core fragments and a heavy atom limit of 10 for terminal (single bond break) fragments and a limit of 12 heavy atoms for core (more than one bond break) fragments.

3. Results and discussion

3.1. Matched molecular pairs and transforms

The identification of matched molecular pairs and their associated transforms is illustrated in Figure 3. In this figure, the two structures in the first column form an example pair where the only difference is a pyridyl (M1) to phenyl (M2) replacement of the central ring. Although this is the only change, there are multiple transforms that represent this pair. The first is a single bond break at the pyrrolidine to generate a methylpicolinamide to methylbenzamide transformation. The second is the pyridyl to phenyl transform that involves two bond breaks. The final is also a two bond break to yield a picolinamide to benzamide transformation. This representation of multiple transforms for the same pair is for the benefit of deriving structure–activity relationships (SAR) from the set of matched molecular pairs. For some endpoints, the simplest pyridyl to phenyl transform may be the best at explaining the underlying SAR, but for other endpoints it may only be the amido substituted pyridyl to phenyl transforms that exhibit a strong SAR signal. Full enumeration of the transforms within the pairs is required to identify these differences. Once all of the matched molecular pairs have been identified in a data set, the activity data is summarized at the transform level to identify SAR.

3.2. Summary pairwise results for data sets

Table 1 shows summary statistics from running PairFinder (our in-house implementation for pair finding) on the human microsomal clearance (HLM), passive membrane permeability (RRCK), P-gp efflux (MDR), and lipophilicity (SF Log *D*) data sets. The number of theoretical compound pairs is calculated as $N(N-1)/2$ where N is the number of distinct compounds in the data set.

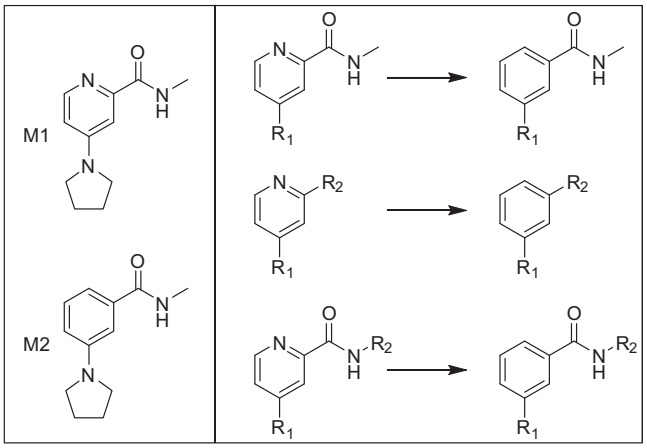


Figure 3. Example of two molecules (M1 and M2) that form a matched molecular pair. The second column contains the three transforms that are enumerated for this pair using the PairFinder algorithm.

The number of possible transforms is higher than the number of distinct compound pairs since a single pair can have multiple transforms (Fig. 3). The number of transforms found trend relatively well with the number of compounds in the data sets. Most of the transforms within a data set occur in only one pair. The number and relative percentages of transforms with occurrences equal to or greater than 2, 5 or 10 are also presented in Table 1. These relative percentages also increase with increasing number of compounds in the data sets. Another interesting trend from Table 1 is the increasing relative percentage of 2-bond and 3-bond break core transforms as N increases. This is due to the increased potential complexity of these fragments relative to single bond break fragments and an increasing likelihood of observing them as the data set size increases.

3.3. Transform examples

Figure 4 shows the top 20 most frequent transforms for HLM, RRCK, MDR, and SF Log *D*. Table 2 contains summary statistics for these transforms including the number of pair examples (n), mean response change (μ), the sample standard deviation of the response change (s), and the matched pairs Student's t-test p-value for the change (p). Also in Table 2 is the percentage of the pairs with an increase, decrease, or minimal change (<20%) in activity (%*I*/*D*/*S*). As expected, most of the top 20 transforms involve very common substitutions in organic molecules (i.e., H → CH_3 (**1**) and CH_3 → CH_2CH_3 (**5**)). There are also the three disubstituted phenyl regioisomer transforms (i.e., o-Ph → m-Ph (**16**), o-Ph → p-Ph (**18**), and m-Ph → p-Ph (**6**)). It is interesting to note that even though many of the transforms listed are statistically significant, most of them have high variability as demonstrated by their high standard deviations relative to their means. For example, the H → OCH_3 (**3**) transform in HLM has a mean change of 10.2 and a SD of 65.4. Also, 35% of the pairs show an increase in activity, while 22% decrease, and 42% show minimal change. This indicates that, although on average, replacement of a proton with a methoxy group results in an increase in microsomal clearance, there are numerous examples where that is not the case.

Despite their large variability, many of these small, simple transforms display very tight additive behavior in their mean response changes. Several examples of this are given in Table 3 for all of the endpoints. The reported Δ values in Table 3 are the differences in activity between the one step transform and the same change but via a two step transform. In the first example, the o-Ph → p-Ph (**18**) transform is compared to the same change via two steps: o-Ph → m-Ph (**16**) and m-Ph → p-Ph (**6**). For HLM, the change in the one step transform (o-Ph → p-Ph) is −24.9 compared to −21.6 for the two step change [(o-Ph → m-Ph = −6.2) + (m-Ph → p-Ph = −15.4)]. Most of the Δ values are low indicating that for these examples, the one step transform can be viewed as a composite of the two step transform. This additive behavior is expected for SF Log *D* as it is the basis for the fragment approach of calculating *c* Log *P*. Thus the lipophilic change from H → Et should approximate the lipophilic change from H → Me plus the change from Me → Et. Lipophilicity changes may also explain the additive behaviors in other endpoints (i.e., HLM and RRCK) in the absence of other mechanisms of action. There are other additive transforms that appear to exist that are not as easy to explain by changes in lipophilicity. One example is the previously discussed o-Ph → p-Ph transform compared to the two step o-Ph → m-Ph and m-Ph → p-Ph transforms (first entry in Table 3) for HLM and RRCK. A significant lipophilicity change is not anticipated for this set of regioisomeric changes, suggesting an alternative perhaps steric explanation for the behavior. For MDR, the near zero Δ values are less informative as most of the individual transforms listed have mean changes near zero.

Table 1
Summary transform information for analyzed data sets

	HLM	RRCK	MDR	SF Log *D*
Compounds	226,348	102,933	74,624	29,998
Theoretical compound pairs	25,616,595,378	5,297,549,778	2,784,333,376	449,925,003
Transforms found	11,973,801	4,338,414	2,633,513	930,892
Unique transforms	7,831,218	2,927,971	1,873,098	761,300
Transforms ($N \geqslant 2$)	1,205,627 (15.4%)	465,718 (15.9%)	271,572 (14.5%)	75,845 (10.0%)
Transforms ($N \geqslant 5$)	217,736 (2.8%)	78,446 (2.7%)	40,866 (1.6%)	8,165 (0.9%)
Transforms ($N \geqslant 10$)	73,118 (0.9%)	23,530 (0.8%)	11,925 (0.5%)	2,228 (0.3%)
Terminal transforms	1,991,936 (25.4%)	924,653 (31.6%)	626,494 (33.4%)	273,103 (35.9%)
Core transforms (R1, R2)	4,355,030 (55.6%)	1,554,563 (53.1%)	968,698 (51.7%)	388,312 (51.0%)
Core transforms (R1,R2,R3)	1,484,252 (19.0%)	448,755 (15.3%)	277,906 (14.8%)	99,885 (13.1%)

Terminal transforms are transforms with single bond break fragments and core transforms are transforms with two or three bond break fragments.

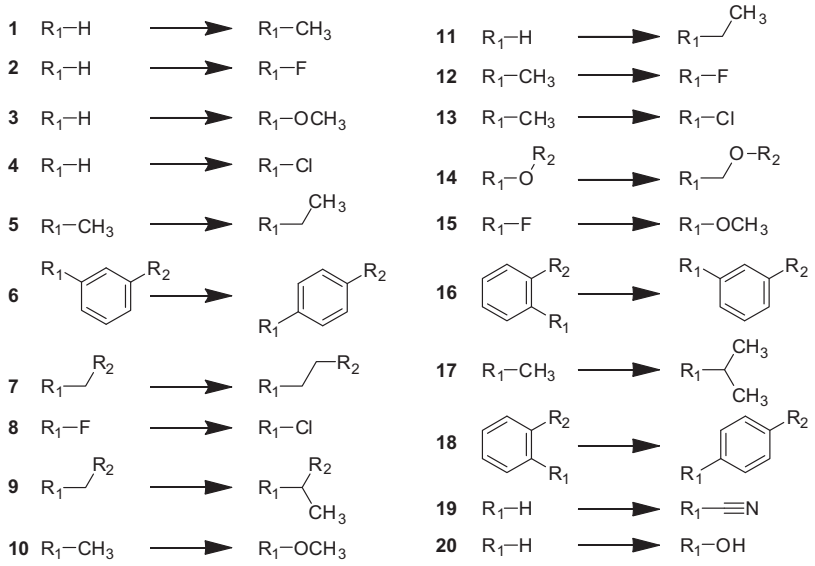


Figure 4. Top 20 most frequently occurring transforms in Pfizer ADME data sets.

3.4. Differences in transform types

A simple calculation of the mean changes in activity is insufficient to fully explain the different types of transformations that appear. The mean activity change assumes that the changes will be more or less constant (with some error) for all pairs. Although this is likely for additive (activity cliff) transforms and bioisosteres, other patterns do exist. For example, there are transforms that demonstrate a proportional change in activity, meaning that the extent of activity change is a function of the starting activity. These transforms would be multiplicative or fold-change transforms. It is important to note that these multiplicative transforms have an additive transform pattern in a log(activity) space. Figure 1 shows some of the different activity patterns that can occur for transforms and these different types are discussed below with examples.

3.4.1. Bioisosteres

The identification of bioisostere transforms is useful for projects whose series may be in appropriate ADME space, but are interested in exploring other aspects of SAR either for the project's primary target activity, off-target activities, or other ADME endpoints. One advantage of the pair finding algorithm is that it finds all pairs without consideration of the underlying change in activity. Consequently, when pairs are summarized by transform, many have little or no associated change in activity. These transforms can be considered bioisosteres for that endpoint. An example is the SF Log *D* bioisostere transform phenyl → 3-methoxyphenyl which contains 48 pairs from the database (Fig. 5(a)). In this figure and the ones to follow, the blue line represents the line of unity and the green line is the total least squares linear regression fit through the points. Note that for this transform the regression line matches the line of unity almost exactly. This is a sign of a bioisostere transform. The pairwise mean change for this transform is −0.02. For validation, the reported Hansch aromatic π value for methoxy is also −0.02.[13] Another bioisostere example is the 3-pyridyl → 4-pyridyl transform for MDR efflux shown in Figure 5(b) consisting of 160 pairs. It has long been known in-house that both 3-pyridyl and 4-pyridyl analogs have a high, and nearly equivalent, propensity to trigger P-gp mediated efflux. Thus, the observation that this transform behaves as a bioisostere for this endpoint is expected.

3.4.2. Additive or activity cliff transforms

Another basic type of transformation is the additive or activity cliff transform. Recently, there has been significant interest in activity cliffs, and what they mean to structure–activity relationships (SAR) and the nature of SAR landscapes.[14–18] Bajorath and Wassermann recently published a pairwise analysis of activity cliffs in compound-protein binding data.[19] Pairwise analysis of our ADME endpoints also reveal transforms that exhibit additive or activity cliff like behavior. One difficulty with activity cliffs however is that it is not always clear where to draw the line between

Table 2
Statistics for top 20 transforms across HLM, RRCK, MDR, and SF Log *D* data sets

Transform	HLM					RRCK					MDR					SF Log *D*				
	n	μ	*s*	*p*	%*I*/*D*/*S*	*n*	μ	*s*	*p*	%*I*/*D*/*S*	*n*	μ	*s*	*p*	%*I*/*D*/*S*	*n*	μ	*s*	*p*	%*I*/*D*/*S*
1	28,181	21.9	61.2	<0.0001	45/12/41	11,461	−0.2	8.8	0.0024	24/28/47	7,269	0.0	6.2	0.7599	20/19/59	2,026	0.32	0.47	<0.0001	21/2/75
2	12,942	2.7	44.8	<0.0001	24/18/56	4,860	−0.5	8.1	0.0001	17/26/55	3,424	0.0	4.6	0.7709	14/16/68	842	0.12	0.38	<0.0001	11/5/83
3	6,816	10.2	65.4	<0.0001	35/22/42	2,844	−1.4	8.0	<0.0001	16/35/48	1,860	1.0	6.0	<0.0001	31/14/54	523	0.02	0.41	0.2610	12/13/74
4	6,645	10.0	61.6	<0.0001	38/19/42	2,137	−3.1	9.4	<0.0001	16/48/35	1,488	0.2	4.4	0.1076	25/16/57	500	0.52	0.39	<0.0001	41/2/57
5	6,007	20.5	50.2	<0.0001	45/9/45	2,169	−1.0	8.1	<0.0001	20/31/48	1,474	0.2	4.8	0.1033	17/15/67	320	0.38	0.36	<0.0001	26/0/72
6	4,710	−15.4	58.4	<0.0001	12/37/49	1,966	−0.1	8.1	0.7840	20/19/60	952	−0.5	3.8	<0.0001	13/19/67	354	0.00	0.30	0.7977	5/5/89
7	4,893	21.2	60.9	<0.0001	43/11/45	1,675	−1.3	8.5	<0.0001	18/32/48	1,017	0.8	6.3	0.0001	26/10/62	291	0.22	0.38	<0.0001	18/3/78
8	4,029	10.0	45.8	<0.0001	37/13/48	1,451	−2.9	8.8	<0.0001	12/48/39	971	0.3	3.2	0.0081	22/11/66	327	0.40	0.27	<0.0001	30/0/68
9	3,659	12.0	49.5	<0.0001	39/13/47	1,531	−0.7	6.8	0.0001	18/23/58	895	0.5	4.9	0.0009	23/13/63	227	0.31	0.38	<0.0001	22/3/74
10	3,103	−8.8	59.4	<0.0001	19/35/45	1,270	0.3	8.0	0.1780	28/25/45	821	0.4	5.8	0.0555	24/22/52	235	−0.21	0.60	<0.0001	6/25/67
11	3,296	44.6	82.3	<0.0001	60/9/29	1,112	−0.8	10.8	0.0160	30/33/35	738	−0.1	5.7	0.5802	26/19/53	168	0.90	0.61	<0.0001	67/0/32
12	2,998	−19.9	56.4	<0.0001	10/45/43	1,215	1.1	8.9	<0.0001	33/18/47	849	−1.0	4.9	<0.0001	9/29/60	220	−0.22	0.37	<0.0001	2/15/81
13	2,937	−8.1	55.2	<0.0001	19/29/51	1,144	−1.3	8.7	<0.0001	19/34/46	704	−0.7	5.6	0.0008	15/21/62	208	0.14	0.41	<0.0001	5/3/90
14	2,870	13.5	59.4	<0.0001	40/15/43	1,019	−1.7	8.1	<0.0001	18/33/48	604	0.4	5.2	0.0957	26/13/59	260	0.22	0.34	<0.0001	29/5/65
15	2,647	12.9	61.8	<0.0001	37/16/45	1,057	−0.6	7.5	0.0068	21/26/52	732	1.5	5.7	<0.0001	36/9/54	207	−0.16	0.43	<0.0001	4/15/79
16	2,817	−6.2	61.3	<0.0001	21/26/52	1,103	−1.4	8.4	<0.0001	16/30/53	556	0.2	4.5	0.2966	18/15/65	149	0.12	0.54	0.0076	13/6/79
17	2,537	28.5	71.2	<0.0001	55/10/33	1,053	−2.7	8.7	<0.0001	23/40/36	756	0.1	6.4	0.6645	31/19/49	156	0.83	0.53	<0.0001	71/0/28
18	2,737	−24.9	73.2	<0.0001	15/42/41	1,068	−1.6	8.0	<0.0001	15/31/52	531	−0.1	4.0	0.6832	21/16/62	127	0.09	0.36	0.0071	14/3/81
19	2,290	−1.4	64.6	0.3114	23/32/43	1,176	−1.3	8.1	<0.0001	19/34/45	709	2.0	7.5	<0.0001	41/13/45	246	−0.28	0.44	<0.0001	4/28/67
20	2,439	−32.0	67.7	<0.0001	11/54/33	1,045	−3.1	11.7	<0.0001	23/47/28	661	4.0	11.0	<0.0001	57/12/30	207	−0.61	0.56	<0.0001	3/43/52

n = number of matched molecular pairs, μ = mean change, *s* = standard deviation, *p* = matched pairs Student's t-test p-value, %*I*/*D*/*S* = %Increase/%Decrease/%Same(+/− 20%).

additive transforms and activity cliff transforms. This will usually involve the user's interpretation/definition of what is a small change and what is a large or cliff-like change.

Table 3 shows that when there are a large number of examples, the mean activity changes for small, common transformations are generally additive, even though these changes have high variance. In addition to these types of transforms, some endpoints have transforms that show a very clear and consistent additive effect without significant variation. For example, the phenyl → p-chlorophenyl transform in SF Log *D* shown in Figure 6(a) contains 45 pairs where almost all show the same change in SF Log *D* values. The mean change is +0.60 and the standard deviation is a relatively low 0.27. This change is very similar to the reported Hansch aromatic π value of +0.71 for a chloro moiety.[13] Figure 6(b) also illustrates a potential additive transformation for RRCK passive permeability whereby replacement of a 3,5-disubstituted-1-methylpyrazole with a 3,5-disubstituted-isoxazole results in an overall increase in permeability. However, this pattern is clearly more variable than the SF Log *D* example (μ = 3.3; *s* = 5.5).

3.4.3. Multiplicative transforms

Another pattern that emerges from the analysis of pairwise activities is the multiplicative or fold-change transform. This occurs when the extent of change for a pair is a function of the beginning structure's activity. One characteristic of a multiplicative transform is that it becomes an additive transform if the activities are logged. Eq. 1 shows a multiplicative transform effect where slope is the slope of the orthogonal best fit line through the data. Taking the log of both sides of Eq. 1 results in Eq. 2; showing that the effect is now additive.

$$\text{SecondActivity} = \text{slope} \times \text{FirstActivity} \quad (1)$$

$$\log(\text{SecondActivity}) = \log(\text{slope}) + \log(\text{FirstActivity}) \quad (2)$$

For the ADME data sets, HLM is the only endpoint which shows a significant number of multiplicative transforms. One example is the 2,5-disubstituted ethyl-linked pyrimidine → 2,5-disubstituted-pyrimidine transform shown in Figure 7(a). To demonstrate the multiplicative/additive relationship described above, the first versus second activity plot for the logged HLM space is also shown in Figure 7(b). For this transform, the slope of the orthogonal best fit line is 2.86 meaning that the removal of the ethyl linker results in a 65% reduction in clearance or going in the opposite direction, an almost three-fold increase in clearance. A second example for HLM is the 2-azabicyclo[2.2.1]heptane → 2-methylpiperidine transform shown in Figure 7(c,d). For this case, breaking the bicyclic bridge results in a 38% reduction in clearance or making the bicyclic bridge results in a 1.6 fold increase in clearance.

3.4.4. Switches

Perhaps the most exciting type of transformation observed in the database is one we have termed a 'switch'. Switch transforms act to essentially turn an activity on or off, much like a light switch. That is to say, regardless of what the starting value of the endpoint is, the transformation results in approximately the same ending value. Note that this is different from an additive transform with a large mean activity change where regardless of starting value, the activity change is the same albeit large. Figure 8(a) shows a well understood switch transformation for HLM: an ethyl ester to a carboxylic acid. Regardless of the starting microsomal clearance value for the ethyl ester compounds (10–325 μL/min/mg), the corresponding carboxylic acid compounds all exhibit a value near 10 μL/min/mg. For HLM clearance, this is not surprising as the carboxylic acid is likely to be deprotonated at pH 7.4 of the assay and therefore limit binding to the lipophilic cytochrome P-450 pockets. Perhaps a more interesting switch like transformation is the replacement of a 2-pyridylmethyl moiety with a 1-methylpyrrolidin-2-one (Fig. 8(b)).

Table 3
Differences in mean activity changes between single step and multistep transform paths showing their additive nature

Transform(s)	No.	HLM		RRCK		MDR		SF Log *D*	
		μ	Δ	μ	Δ	μ	Δ	μ	Δ
o-Ph → p-Ph	**18**	−24.9	−3.35	−1.6	−0.1	−0.1	0.2	−0.09	−0.03
(o-Ph → m-Ph) + (m-Ph → p-Ph)	(**6** + **16**)	−21.6		−1.5		−0.3		0.12	
H → Et	**11**	44.6	2.2	−0.8	0.4	−0.1	−0.3	0.90	0.20
(H → Me) + (Me → Et)	(**1** + **5**)	42.4		−1.2		0.2		0.70	
H → Cl	**4**	10.0	−2.7	−3.1	0.3	0.2	−0.1	0.52	0.00
(H → F) + (F → Cl)	(**2** + **8**)	12.7		−3.4		0.3		0.52	
H → OMe	**3**	10.2	−2.9	−1.4	1.5	1.0	0.6	0.02	−0.09
(H → Me) + (Me → OMe)	(**1** + **10**)	13.1		0.1		0.4		0.11	

μ is the mean activity change or sum of the mean activity changes for the transform(s). Δ is the difference in activity changes for the different paths.

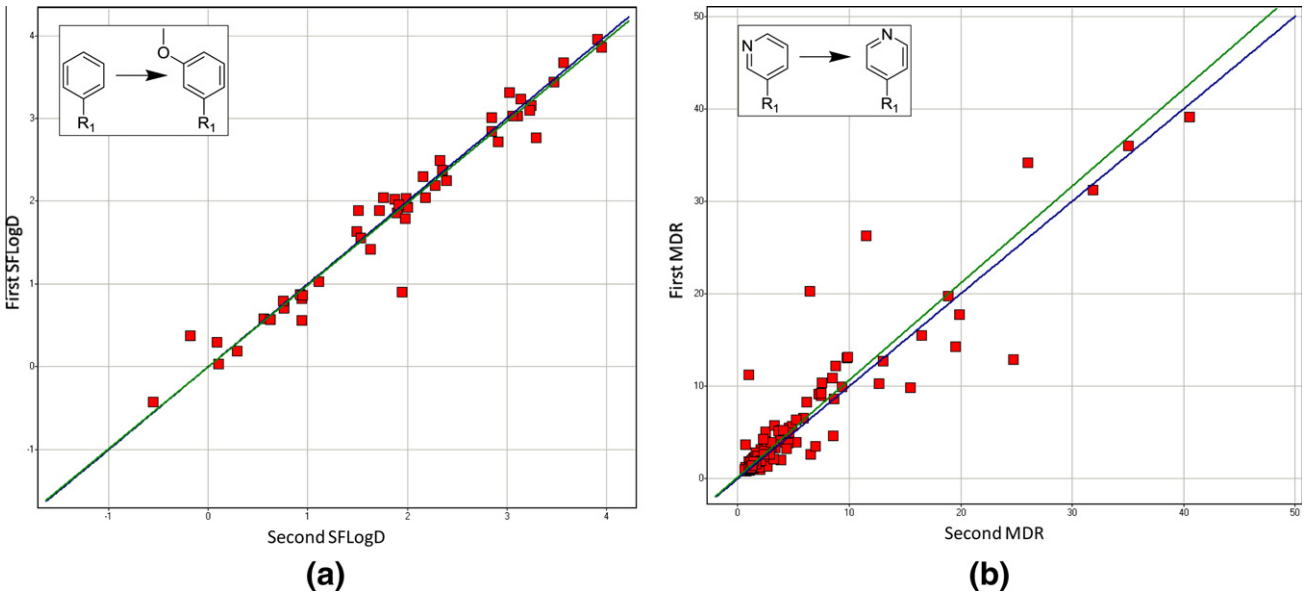


Figure 5. Plots of first versus second pairwise activities for the (a) SF Log *D* phenyl → 3-methoxyphenyl (48 pairs) and (b) MDR efflux 3-pyridyl → 4-pyridyl (160 pairs) bioisostere transforms. Dark blue line is the line of unity and the green line is the total least squares linear regression fit.

Similar to the ester to acid example, regardless of the clearance of the starting pyridyl analog, conversion to the pyrrolidinone results in a decrease in HLM clearance to a similar value, a more moderate 50 μL/min/mg. Note that this transformation retains the hydrogen bonding potential of the initial pyridine while eliminating a potential clearance liability. Of particular interest is identifying transformations with unanticipated switch like behavior. An example is the hydrogen → 4-piperidine transform (Fig. 8(c)) where replacement of a proton with a 4-piperidyl group results in HLM clearance values of ~20 μL/min/mg for all 72 examples, regardless of the starting clearance values for the hydrogen analogs. Switches are also present in the MDR efflux ratio data set. Two examples are highlighted: the 3-pyridyl → phenyl transform (Fig. 8(d)) with 193 pairs and a two-bond break transform involving sulfone → ether (Fig. 8(e)) with 75 pairs. In both of these examples, the transformation results in an elimination of MDR efflux activity. For RRCK, switches are less common, but one example is the phenyl → 4-imidazole switch transform shown in Figure 8(f) where the transformation acts to turn off passive permeability.

3.5. Switch directionality

One interesting observation about switches is their directionality. For other types of transforms, if the A → B transform increases an activity, it is expected that the B → A transform will decrease the activity in an equal way. This is not necessarily true for switches. As illustrated in Figure 8(d), the 3-pyridyl → phenyl switch transform for MDR efflux, almost all of the compounds that contain the phenyl ring have a relatively low MDR efflux ratio between 2 and 5. This implies that if a compound contains a 3-pyridyl group and it is changed to a phenyl group, the result will be a lower MDR efflux ratio. However, the 3-pyridyl compounds have efflux ratios ranging from 2 to 45 with many in the 2–5 range. As a result, if a compound containing a phenyl fragment is transformed to a 3-pyridyl fragment, an increase in MDR efflux ratio is not assured. Therefore, when dealing with switch transforms there must be careful consideration of the directionality of the switch and the likelihood of seeing a meaningful change versus no change at all.

3.6. Frequency of different transform types

To gain a clear understanding of how often bioisostere, additive, multiplicative and switch transforms occur across the different data sets, a set of filters was developed that can be applied to the transform statistics. A bioisostere transform has two theoretical requirements: (1) the slope of the total least-squares fit between activity one and activity two should be near 1.0 (see Fig. 1), and (2) the average activity change should be near zero. Our filter for

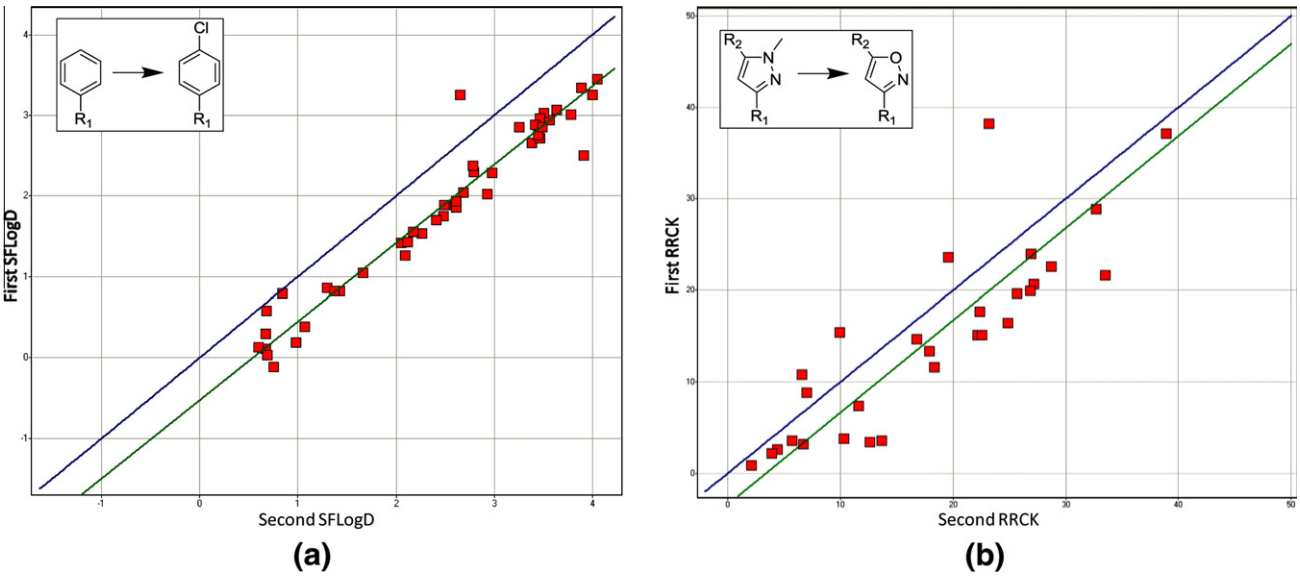


Figure 6. Plots of first versus second pairwise activities for the (a) SF Log *D* phenyl → *p*-chlorophenyl (45 pairs) and (b) RRCK 3,5-disubstituted-1-methylpyrazole → 3, 5-disubstituted-isoxazole (29 pairs) additive transforms. Dark blue line is the line of unity and the green line is the total least squares linear regression fit.

a bioisostere requires a slope between 0.8 and 1.25, and a mean activity change less than one half the overall mean activity change for all transforms. Additive transforms must also meet the first criteria for a bioisostere, but the mean activity change should be greater than zero and the overall error in the change should be small. Our requirement for an additive transform is a slope between 0.8 and 1.25 and a mean activity change greater than twice its standard deviation. Since multiplicative transforms are additive in a logged activity space, we use the additive rules above in the logged space to identify them. For switch transforms, the absolute slope is required to be greater than 5.0 or less than 0.2, the mean activity change has to be greater than twice the overall mean average activity change for all transforms, and the activity one to activity two standard deviation ratio has to be greater than 2 or less than 0.5. An on switch is a switch where the mean activity transformed to is *higher* than the mean activity transformed from. An off switch is a switch where the mean activity transformed to is *lower* than the mean activity transformed from. Note that the above criteria are arbitrary, relatively conservative, and not the only way to identify the patterns shown in Figure 1. There are other transforms that after visual inspection could be classified as bioisosteres, additive, multiplicative, or switch transforms, but are not identified using these criteria because of outliers, general activity variability or other factors. Nonetheless, these filters give a good sense of the likelihood and relative abundance of the different types of transforms in these endpoints.

Table 4 shows the results of applying these filters to the ADME data sets and there are a couple of notable patterns. First, all of the assays have a reasonably high proportion of bioisostere transforms. SF Log *D* has the highest at 33% with MDR efflux the lowest at 17%. For additive transforms, SF Log *D* is the only assay with an appreciably large number at 22.4%. HLM is the only assay that shows a significant number of multiplicative transforms with 522 or 0.71%. There are significantly fewer additive or multiplicative transforms in the other assays.

On switches are extremely rare across all four endpoints (0.02–0.09%) which is also not unexpected. It is unlikely that a small molecular transformation would result in an increase to a constant activity unless the change was associated with the limit of detection for the assay. Upon closer inspection of the few on switches, it does appear that many are either statistical outliers or involve very specific chemical series. On the other hand, off switches occur more frequently, especially in the HLM and MDR efflux data sets. MDR efflux has the highest frequency of off switches at 6.7% while HLM has 4.3%. RRCK and SF Log *D* have a very small percentage of off switches at 0.15% and 0.04%, respectively.

3.7. Switches and molecular recognition

The existence of transforms that behave like switches may suggest that one of the fragments is somehow involved in a specific molecular recognition event. The relative abundance of switches also supports this possibility. The MDR efflux assay measures the effect of the P-glycoprotein (P-gp) efflux transporter. P-gp interacts with a wide variety of substrates and would therefore be expected to have a large number and variety of potential molecular recognition fragments. Table 4 shows it has the highest proportion of switches at 6.7%. High HLM clearance requires that the molecular substrate in question binds to one of several different cytochrome P-450 enzymes. Once bound, it must also be able to form specific interactions between a metabolically reactive fragment in the molecule and the heme cofactor. Since this specific molecular interaction event is a requirement, it would be expected to have a high proportion of switch transformations as well and it has the second highest percentage of 4.3%. The measured SF Log *D* is not expected to be a function of specific molecular recognition events and would therefore have few if any switch transforms. Indeed, only one transform that meets the switch criteria was found in the SF Log *D* database. Finally, RRCK passive permeability does not involve any known specific molecule–protein recognition processes, but does involve molecular interaction of the molecule and cellular membrane which could be affected by some fragment types especially those that change the charge state of the molecule. It has a low percentage of switch transforms at 0.15%.

3.8. PairTransformer

In a typical analog design session by project teams, the efficient and effective generation of ideas and solutions for solving the myriad of issues which prevent a particular compound from becoming a drug candidate is paramount. It is in this environment where pairwise analysis, with its ability to extract (tacit) knowledge and suggest chemical transformations, may provide its most significant utility. In order to apply the pairwise information to new com-

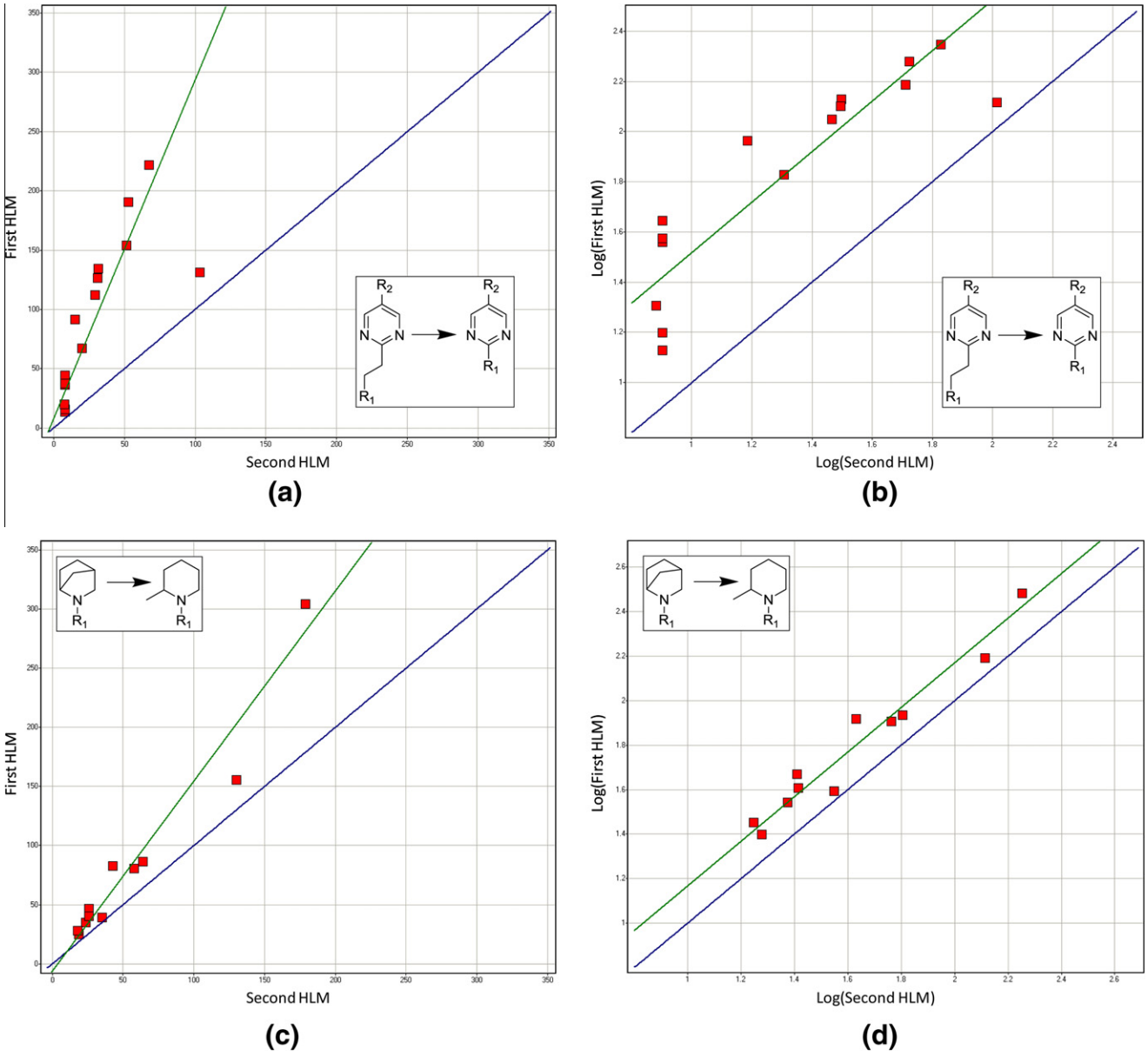


Figure 7. Plots of first versus second pairwise activities for the HLM (a) 2,5-disubstituted ethyl-linked pyrimidine → 2,5-pyrimidine (15 pairs) and (c) 2-azabicyclo[2.2.1]heptane → 2-methylpiperidine (11 pairs) multiplicative transforms. Panels (b) and (d) show their additive pattern in the logged activity plots. Dark blue line is the line of unity and the green line is the total least squares linear regression fit.

pounds and projects, an application called PairTransformer was developed. This tool was written in a modular fashion for execution from our in-house compound analysis tool, PCAT.[11] The interface allows an end-user to select a portion of the query molecule to be searched against the pairwise database. The tool then enumerates transformed products using the transforms found in the database and the original query molecule. The user also has the option to indicate how many pairs must exist in the database to constitute a meaningful transform since the confidence of an activity change associated with the suggested chemical transformation increases as the number of pair examples increases.

Two distinct but complementary use cases of this application are presented. The first use case demonstrates a typical design workflow involving a search of the pairwise database for substituent replacements to improve an ADME endpoint of interest. The second use case demonstrates a search of the database to assess the impact of a structural change across multiple ADME endpoints. PairTransformer returns the following information for each transformed product it generates: the predicted response for the new structure using our in-house in silico ADME statistical models, the number of pairs found in the database, the mean activity change for that transform, the standard deviation of the change, the matched pairs Student's t-test for the change, and the percentage of pairs with increasing, decreasing, and minimal response change (defined as less than 20% for the ADME endpoints).

Use case 1: Identify replacements for the cyclopropyl moiety in JAK inhibitors that decrease microsomal clearance

In this particular example, the cyclopropyl analog of a recently disclosed JAK inhibitor series[20] was selected as a test compound and the goal was to identify potential cyclopropyl replacements which would result in a reduction in HLM clearance. Figure 9 displays the results of the PairTransformer search for this query. The first row of the results table shows the input structure and each subsequent row displays a simplified schematic of the transformed

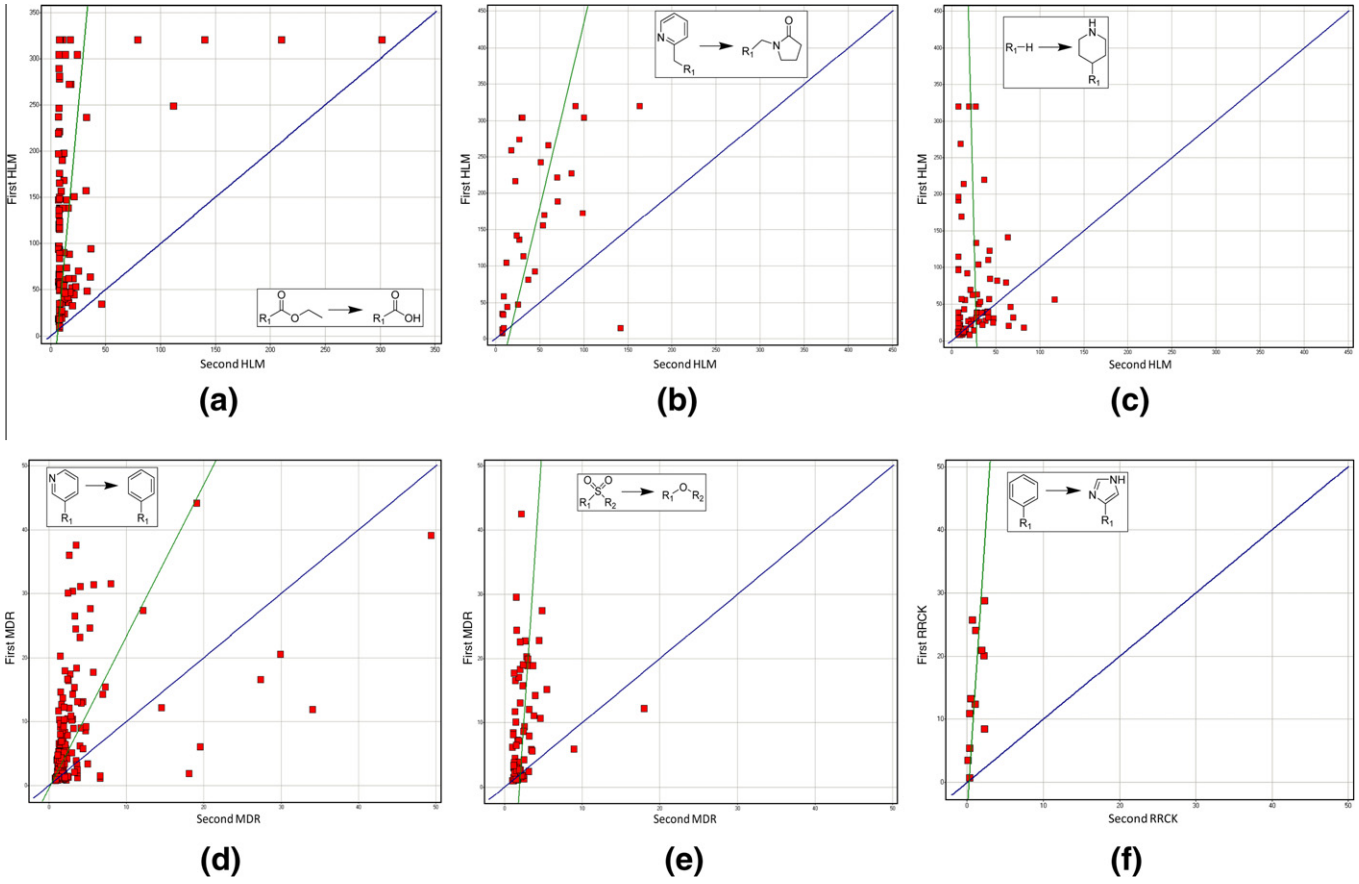


Figure 8. Plots of first versus second pairwise activities for the (a) HLM ethylester → carboxylic acid (127 pairs), (b) HLM 2-pyridylmethyl → 1-methylpyrrolidin-2-one (33 pairs), (c) HLM hydrogen → 4-piperidine (72 pairs), (d) MDR efflux 3-pyridyl → phenyl (193 pairs), (e) MDR efflux sulfone → ether (75 pairs), and (f) RRCK phenyl → 4-imidazole switch transforms. Dark blue line is the line of unity and the green line is the total least squares linear regression fit. Note the switch like behavior 'turning off' the respective activities.

Table 4
Number and relative percentage of bioisostere, additive, multiplicative, and switch transforms in the ADME data sets

Assay	Bioisostere	Additive	Multiplicative	On switch	Off switch
HLM	14,035 (19.2%)	19 (0.03%)	522 (0.71%)	11 (0.02%)	3,112 (4.3%)
RRCK	4,844 (20.6%)	67 (0.28%)	63 (0.27%)	16 (0.07%)	346 (0.15%)
MDR	2,025 (17.0%)	3 (0.03%)	30 (0.25%)	2 (0.02%)	800 (6.7%)
SF Log *D*	739 (33.2%)	499 (22.4%)	N/A	2 (0.09%)	1 (0.04%)

product and the transform applied. The experimental human liver microsomal (HLM) clearance values of these analogs have been published and are shown in the last column of Figure 9 (Experimental CL).[20] For the cyclopropyl analog search, five transformations are highlighted. For four of the five examples, PairTransformer suggested replacements for the cyclopropyl ring that would result in a decrease in microsomal clearance. The presence of multiple pair examples and agreement by the in silico model predictions (third column in Fig. 9 headed by cHLM_01_CLIA) provide increased confidence that these transforms will yield the desired effect. For these four examples, the predictions were confirmed by experimental data. For one of the transforms, replacement of the cyclopropyl ring with a thiophene (row two in Fig. 9), the pairwise transform data suggest that there would be an increase in HLM clearance (mean change = 67 μL/min/mg and 75% of pairs show an increase). However, the number of pair examples in this set is small (NumExamples = 4). The in silico model prediction did not agree with the pairwise result and calculated a slight decrease for the transformed structure. The actual experimental data shows a decrease in microsomal clearance for this transformation (70 → 46 μL/min/mg). This table highlights the various endpoints that need to be considered to assess the potential effect of a pairwise transformation. While the mean change (Delta_Mean) of the members of a transform can give a quantitative signal of the effect, the distribution of the data in less clear and the magnitude of the effect can be skewed by extreme values in the set. Conversely, while the relative distribution of changes (PctInc, PctDec, PctSame) can give an indication of opposite effects that may be present for the same transformation, the magnitude of the effect is not defined. The numbers of examples in the transform (NumExamples) should also be considered as higher numbers may provide greater confidence in the overall direction of the effect. Ultimately, those transforms with numerous matched molecular pair examples that also show a large mean change supported by an appropriate change distribution have the highest probability to produce the desired effect experimentally.

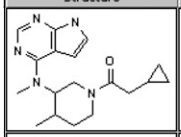
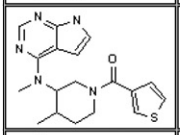
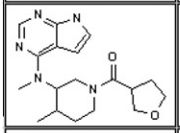
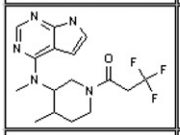
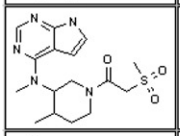
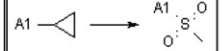
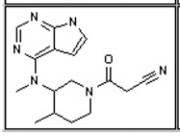
Structure	Transform_Smiles_stx	cHLMG_01_CLIA	NumExamples	Delta_Mean	Delta_StdDev	Delta_PValue	PctInc	PctDec	PctSame	Experimental CL
		**98.1**	**0**	**0.0**	**0.0**	**0.0**				**70**
		83.4	4	67.1359	54.6994	0.0912949	75%	0%	25%	46
		40.0	86	-31.6044	53.9289	5.15443e-07	10%	62%	26%	11
		24.6	414	-11.4401	62.3663	0.000216278	15%	36%	47%	16
		8.02	22	-43.3355	73.3647	0.0114605	4%	68%	27%	8
		8.43	77	-46.1378	73.9337	5.40252e-07	9%	67%	23%	10

Figure 9. Selected PairTransformer results from a search of the cyclopropyl moiety in the JAK compound (top row) for transforms that reduce microsomal clearance. The last column shows actual experimental HLM values.

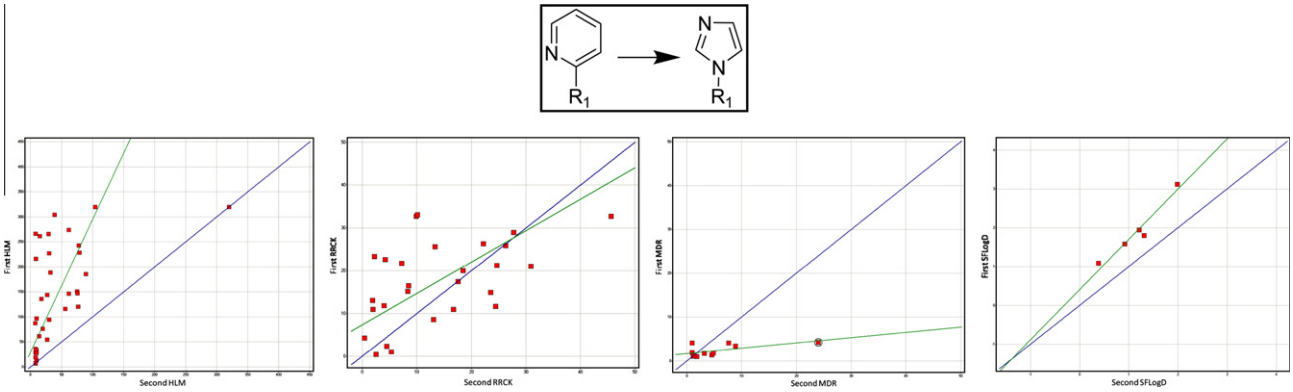


Figure 10. Results for search of the 2-pyridyl → 1-imidazole transform across all ADME endpoints.

Use case 2: examine the overall ADME effect of transforming 2-pyridyl to 1-substituted imidazole (Fig. 10)

In this example, the impact of replacing a pyridyl with an imidazole across multiple ADME endpoints is assessed. To perform this experiment, a search for each ADME endpoint of interest is invoked. For microsomal clearance, 35 pairs exist, with the majority indicating a lower clearance for the imidazole analog. This example accentuates the complexity of the system that these patterns are trying to decipher. The potential of 1-substituted imidazoles to inhibit cytochrome P450s is well known[21] and thus, the observed decrease in microsomal clearance in this transform is likely a consequence of this inhibition as opposed to changing the intrinsic metabolic stability of the compound. The pattern for the RRCK change is not well defined and a systematic change in permeability is not apparent. Conversely, there are 11 pair examples in the MDR efflux database, with most pointing to an increase in efflux ratio for the imidazole analog. Only five pair examples exist in the SF Log *D* database, but in all cases the imidazole analog is less lipophilic with an average SF Log *D* change of −0.7. The net result is a transform that would reduce the apparent clearance and lipophilicity, potentially increase the P-gp efflux (though the limited range of the dataset precludes a definitive assessment), and cause an unknown effect on permeability. However, in this case, the mode by which the apparent reduction in clearance is achieved may not be desirable.

4. Conclusion

We have described PairFinder, a process for generating a comprehensive catalog of matched molecular pairs with their corresponding ADME data. In essence, this represents a codification of the tacit or soft knowledge embedded in the activities of the

compounds. PairTransformer allows for efficient and effective search of these pairwise databases for idea generation with the ability to propose solutions for a variety of questions. These ideas and solutions are based on transformations which other past or current teams have employed for their series and that collectively resulted in a desired change in activity. Statistical analyses of these transforms coupled with in-house in silico model prediction capabilities provide context and confidence in the activity of the proposed structures. A distinctly different but complementary use of this information is to assess the impact of a particular transformation across a series of ADME endpoints. This second approach is critical as consideration of multiparametric optimization is essential for effective development of potential drug candidates.

To further extract the embedded knowledge in the structure and activities of the matched molecular pairs, a systematic analysis of the activity patterns has revealed that there are a variety of transform types that exhibit characteristic, but complex patterns. Because of this, it is difficult to capture the exact nature of the molecular transforms using simple summary statistics. The transform types include bioisosteres, which do not result in any significant change in response; additive transforms, which change a response by a consistent amount; cliff transforms, which behave similarly to additive transforms but with greater magnitude; multiplicative or fold-change transforms where the magnitude of change is proportional to the beginning activity; and finally, switch transforms that act to 'turn on' or 'turn off' the response. Switches also tend to be unidirectional meaning that unlike additive transforms, if a switch can be used to turn off a response, the inverse of that switch will not necessarily turn on the response. The size of the data sets analyzed has resulted in a tremendous amount of information with thousands of interesting transforms for each one. Detailed analysis of each of the endpoints is ongoing and will be the subject of future communications. The combination of these two tools and our processes to keep the information current via automated regular updates represents our efforts to capture, mine and extract organizational ADME SAR knowledge.

Acknowledgments

This research was sponsored by Pfizer Inc. The authors gratefully acknowledge Mark Gardner, James Mills, Jared Milbank, Hualin Xi, and Hao Sun for helpful discussions and analysis of preliminary results.

References and notes

1. Nonaka, I.; von Krogh, G. *Organ. Sci.* **2009**, *20*, 635.
2. Sheridan, R. P.; Hunt, P.; Culberson, J. C. *J. Chem. Inf. Model.* **2006**, *46*, 180.
3. Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. *J. Med. Chem.* **2006**, *49*, 6672.
4. Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. *Bioorg. Med. Chem.* **2009**, *17*, 5906.
5. Lewis, M. L.; Cucurull-Sanchez, L. *J. Comput. Aided Mol. Des.* **2009**, *23*, 97.
6. Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. *J. Chem. Inf. Model.* **2010**, *50*, 1350.
7. Cucurull-Sanchez, L. *J. Comput. Aided Mol. Des.* **2010**, *24*, 449.
8. Hussain, J.; Rea, C. *J. Chem. Inf. Model.* **2010**, *50*, 339.
9. Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W.; Macdonald, S. J. *J. Chem. Inf. Model.* **2010**, *50*, 1872.
10. OpenEye Scientific Software, Santa Fe, NM, 87507. 2010.
11. *PCAT*. Pfizer Compound Analysis Tool, version 4.1, Pfizer Global Research & Development, Cambridge, MA, 2010; PCAT is a tool for clustering, organizing, and visualizing molecules with their associated properties and biological activities.
12. Hop, C. E.; Cole, M. J.; Davidson, R. E.; Duignan, D. B.; Federico, J.; Janiszewski, J. S.; Jenkins, K.; Krueger, S.; Lebowitz, R.; Liston, T. E.; Mitchell, W.; Snyder, M.; Steyn, S. J.; Soglia, J. R.; Taylor, C.; Troutman, M. D.; Umland, J.; West, M.; Whalen, K. M.; Zelesky, V.; Zhao, S. X. *Curr. Drug Metab.* **2008**, *9*, 847.
13. Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 525.
14. Peltason, L.; Bajorath, J. *J. Med. Chem.* **2007**, *50*, 5571.
15. Maggiora, G. M. *J. Chem. Inf. Model.* **2006**, 46.
16. Guha, R.; Van Drie, J. H. *J. Chem. Inf. Model.* **2008**, *48*, 646.
17. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. *J. Med. Chem.* **2008**, *51*, 6075.
18. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. *Drug Discovery Today* **2009**, *14*, 698.
19. Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model.* **2010**, *50*, 1248.
20. Flanagan, M. E.; Blumenkopf, T. A.; Brissette, W. H.; Brown, M. F.; Casavant, J. M.; Shang-Poa, C.; Doty, J. L.; Elliott, E. A.; Fisher, M. B.; Hines, M.; Kent, C.; Kudlacz, E. M.; Lillie, B. M.; Magnuson, K. S.; McCurdy, S. P.; Munchhof, M. J.; Perry, B. D.; Sawyer, P. S.; Strelevitz, T. J.; Subramanyam, C.; Sun, J.; Whipple, D. A.; Changelian, P. S. *J. Med. Chem.* **2010**, *53*, 8468.
21. Franklin, M. R.; Constance, J. E. *Drug Metab. Rev.* **2007**, *39*, 309.