Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) rules from Cross-company Matched Molecular Pairs Analysis (MMPA)

Journal:	*Journal of Medicinal Chemistry*
Manuscript ID	jm-2017-00935p.R2
Manuscript Type:	Perspective
Date Submitted by the Author:	12-Sep-2017
Complete List of Authors:	Kramer, Christian; F. Hoffmann-La Roche AG, Research and Development Division Ting, Attilla; AstraZeneca plc Zheng, Hao; Genentech Inc Hert, Jérôme; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Schindler, Torsten; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Stahl, Martin; F Hoffmann-La Roche AG Research and Development Division, Roche Innovation Center Basel Robb, Graeme; AstraZeneca plc Crawford, James; Genentech Inc Blaney, Jeff; Genentech Inc Montague, Shane; MedChemica Limited Leach, Andrew; MedChemica Limited Dossetter, Alexander; MedChemica Limited Griffen, Edward; MedChemica Limited

Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) rules from Cross-company Matched Molecular Pairs Analysis (MMPA)

Christian Kramer[*, §], Attilla Ting[*, •], Hao Zheng[*, ‡], Jerome Hert[§], Torsten Schindler[§], Martin Stahl[§], Graeme Robb[•], James J. Crawford[‡], Jeff Blaney[‡], Shane Montague[†], Andrew G. Leach[†], Al. G. Dossetter[†], Ed J. Griffen[*, †]

[§]Roche Pharma Research and Early Development, Roche Innovation Center Basel, Switzerland, [‡]Genentech Inc, 1 DNA Way, South San Francisco, CA 94080, [•]AstraZeneca plc, Milton Rd, Milton, Cambridge, CB4 0FZ, [†]MedChemica Ltd, Biohub Alderley Park, Macclesfield. Cheshire SK10 4TG.

Abstract

The first large scale analysis of in vitro Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) data shared across multiple major Pharma has been performed. Using advanced matched molecular pair analysis (MMPA), we combined data from three pharmaceutical companies and generated ADMET rules, avoiding the need to disclose the full chemical structures. On top of the very large exchange of knowledge, all companies involved synergistically gained approximately 20% more rules from the shared transformations. There is good quantitative agreement between the rules based on shared data compared to both individual companies' rules and rules published in the literature. Known correlations between logD, solubility, in vitro clearance and plasma protein binding also hold in transformation space, but there are also interesting exceptions. Data pools such as this allow focusing on particular functional groups and characterizing their ADMET profile. Finally the role of a corpus of robustly tested medicinal chemistry knowledge in the training of medicinal chemistry is discussed.

Keywords

ADMET, MMPA, matched molecular pair analysis, data sharing, data mining, big data, unsupervised learning, precompetitive research, medicinal chemistry knowledge.

Introduction

Medicinal chemistry is a discipline at the borderline of science and technology that depends on experience, intuition and knowledge of rules that govern the space of medicinal chemistry practice. As such, successful medicinal chemistry optimization has routinely drawn on knowledge gained from past experimental observations. Thus it is fair to assume that the more data are available (everything else remaining equal), the more knowledge can be gained.

While more data is desirable, the strategic necessities in a patent-based competitive environment allow the public sharing of only a small proportion all of the information generated inside pharmaceutical companies. In particular, the composition of chemical matter with its associated biological properties is a cornerstone of pharmaceutical patent filing. Therefore parties interested in protecting their research must exercise great care in the chemical structures they disclose either in a patent or in academic publications not to weaken any patent position. It is therefore unsurprising that only a small proportion of chemical structures synthesized and their assay data tested are ever made public. As medicinal chemistry knowledge is the understanding of the relationships between chemical structure and biological properties, this creates a conflict between gaining knowledge and maintaining intellectual property.

In order to increase our knowledge space, we set out on an experiment to combine absorption, distribution, metabolism, and toxicity (ADMET) data from three different large pharmaceutical companies (AstraZeneca, Genentech, and Roche; Genentech research and Roche research operate independently) and mine the pooled data for relevant medicinal chemistry knowledge. MedChemica was used as the intermediary to combine rules. In particular, we were interested in whether we could merge ADMET knowledge between the companies, and whether we could synergistically increase our medicinal chemistry knowledge based upon the joint information pool.

A solution to sharing large amounts of medicinal chemistry data between companies in the patent-based competitive environment is to use matched molecular pair analysis (MMPA),[1–7] since the original structures of the compounds cannot be calculated back if only the part of the structure which is involved in the transformation is shared.[8] MMPA extracts transformations that link a pair of compounds. For all pairs that are linked by the same transformation, MMPA calculates aggregate statistics of the property differences to derive a rule for the given transformation. In this work we use the term “rule” to indicate a “virtual” chemical transformation leading to a change in a physical chemistry or biological endpoint where there are enough examples to support an assertion as to how this change influences certain given molecular properties. Very rarely, there are transformations where all known pairs change property in the same direction, in particular if there are many pairs. Nevertheless, even in the presence of exceptions, knowing the most common effects is useful to guide whether or not a compound shall be made. The rules therefore capture variability, which is a crucial element in medicinal chemistry decision making. Obviously more than one pair is needed for a rule to be more than an anecdote; the statistical approach used here suggests an absolute minimum of six pairs should be considered in order to form a rule. One of the most important attributes of MMPA is its closeness to medicinal chemistry reasoning, where talk of “adding a fluorine to block metabolism” or “adding a solubilizing group” are common descriptions of underlying thinking. Occam’s razor suggests that the simplest cause for a change in a property is the single change in structure. Neighboring groups within the molecules involved may also play a causal role, hence, the local chemical environment is incorporated into the encoding of the structural change and should allow these effects to be partitioned and examined separately.[9] The only remaining explanations are long range effects or non-additivity of chemical changes. Kramer et al showed that SAR non-additivity may be masked due to experimental uncertainty in many cases, and in other cases depends on the target and binding mode.[10]

There have been significant studies, already published, describing the application of MMPA to ADMET datasets from individual pharmaceutical companies.[2,11,12] Most of these have been supervised analyses, where a particular question is asked such as “what effect does hydrogen replacement on an aryl ring have on ADMET properties”[13,14] or “what are the effects of substituting different heterocycles for phenyls".[12] In contrast, here we attempt to characterize the entire ADMET rule space.

Having developed the technology to merge data between companies (details can be found in supporting information), the next question becomes: "How many extra rules do we gain and is there any degradation in knowledge by merging data between companies?" It is relatively obvious that by aggregating pairs from different companies the statistics for some rarer transformations will become significant because there are more pairs. However, this gain may be obfuscated by noise introduced due to mixing data from assays with the same endpoint but run under different conditions.[15] A major concern is assay comparability: although two companies may formally measure the same ADMET attribute, if the assays used are different the outcomes may be different. The advantage of MMPA here is that two compounds are only compared if they have been measured by the same operator using the same method and endpoint. The difference in an attribute between two compounds is then calculated. If the assays are merely systematically displaced against other, then the systematic displacement is automatically factored out by using differences rather than absolute values. It is then possible to test if pairs can be aggregated across different assays and rules identified. To validate this, we (a) compare the rules we found by aggregating pair sets across companies to rules obtained from one company only and (b) to rules previously published.

High-quality analysis from MMPA requires large amounts of data, and we will show that more data is clearly better with brief examples of analyses that can be made using such a massive dataset. We will illustrate the types of analyses that can be done having a rule pool as large as ours. We initially present general analyses of correlations of solubility, clearance, and PPB rules with logD (pH 7.4) rules. As rules representing general knowledge are always accompanied by exceptions, we highlight a few less-intuitive medicinal chemistry rules that counter the general trends. Finally, we give a taste of how such a huge set of rules can be used to characterize individual functional groups, a statistical type of analysis which only becomes possible once large MMP datasets are available. It is not our intent to hypothesize the cause of a particular quoted rule being significant as a thorough exploration of the rules has been beyond the scope of this work, and value can be obtained from their knowledge without complete clarity as to their causation.

Methods

Data extraction

Data needs to be transformed into a normally distributed variable before MMPA *i.e.* pIC_{50}, $\log_{10}$ (molar solubility), $\log_{10}$(intrinsic Clearance, (CL_{int})) for in vitro clearance assays, log(free/bound) for protein binding etc. Out of range flags were attached where appropriate. If there were multiple measurements in the same assay for the same compound then the data were aggregated using the following scheme:

1. Where there are no qualifiers / out of range flags:

   - aggregate as median logged value – as this is more representative of true value than mean, and less sensitive to outliers

2. Where there are some qualified data present:

   - take the median of the data excluding qualified data:

     For example: 3, 6, >8, >10, <3, 5, 6: exclude >8, >10, <3, the median(3,6,5,6)=5.5

3. If there are only qualified data present:

- if there is a mixture of > and < then discard the data
- if the qualifiers are all < then use the minimum value
- if the qualifiers are all > then use the maximum value.

We used $\log D_{7.4}$, solubility, in vitro microsomal and hepatocyte clearance (human, rat, mouse, cynomolgus monkey, dog), Madin Darby Canine Kidney cells (MDCK cells) permeability (A-B, B–A and efflux), cytochrome P450 inhibition (2C9, 2D6 , 3A4 , 2C19 , 1A2), NaV 1.5 and human ether-a-go-go related gene product (hERG) ion channel inhibition, glutathione stability, and PPB (human, rat ,mouse, cynomolgus monkey, dog) assays. We have not addressed explicitly the propagation of experimental errors, or the heteroscedasticity of assay data.[16] In keeping with a "Big Data" paradigm we have chosen to include as much measured data as possible and to use non-parametric statistics to generate inferences that may be more resistant to the effects of noise.

Matched Molecular Pair (MMP) analysis

Compound structures were extracted as SMILES, desalted and had their charge corrected to the neutral form by applying SMIRKS using the Openeye OEChem toolkit. A set of SMIRKS were applied to recursively standardize tautomers to a common form. These were not intended to be completely accurate tautomeric forms for a given physical state, but to provide a unique canonical standard. For chiral compounds, racemates were excluded from the analysis. No further filters were applied to the inclusion of compounds.

An MMP analysis platform, MCPairs, was used that combines the benefits of the Hussain and Rea algorithm[17] with those of the WizePairZ algorithm.[18] Compounds were fragmented on all noncyclic single bonds except amide (C-N) and sulfonamide (S-N) bonds. Fragmentations of up to three cuts were created. Pairs were matched based on a joint constant part, and the non-identical parts were taken as the transformations. The transformations were recorded as SMIRKS with explicit hydrogens to ensure correct aromaticity and tautomerization assignment. For a matched pair change, the local chemical structures up to four bonds from the point of change were captured, which we classify as the "environment" in the same manner as in WizePairZ. Note that within this formalism, two compounds can be linked by several transformations. Pairs were only considered if the part encoded in the SMIRKS did not make up more than 45% of the molecule and the number of non-hydrogen atoms in the changing part of the transformation was 16 or fewer. Pairs were not used if the measured activities for both compounds have a qualifier of the same direction, for example both compounds were measured as > X. An example for a pair is given in Figure 1. More examples can be found in the supporting information Figure S1.

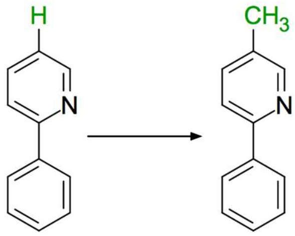
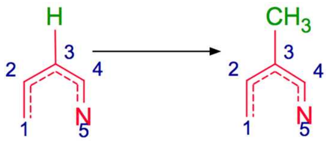
Molecule Pair	 Molecule A → Molecule B
3 bond environment	 H 2 3 4 1 N 5 CH3 2 3 4 1 N 5 [c:1]([H])[c:2]([H])[c:3]([H])[c:4]([H])[n:5]>> [c:1]([H])[c:2]([H])[c:3]([c:4]([H])[n:5])[C]([H])([H])([H])

Figure 1: Functional group transformation with environment specification of 3 bonds as chemical sketch and as SMIRKS notation (examples for other environments can be found in the supporting information). The green groups are those being changed in the transformation, the red portions of the transformations show the increasing level of environment specification increasing in all directions from the point of change. The blue atom numbers correspond to the atom mappings in the SMIRKS transformation encoding.

Software

Data extraction, manipulation and analysis were performed using scripts written in Python[19] using the Openeye cheminformatics toolkit[20] for chemical structure manipulation. MySQL[21] was used for database infrastructure and statistical routines were implemented in R[22] using the rpy2[23] interface to Python.

Rule merging/ calibration

MMP identification was carried out within each contributing company behind their firewalls. Data tables of the transformations with the particular biological or physical chemistry delta values and associated calculated properties were then supplied to MedChemica as a neutral third party. These data tables excluded the original compound structures. The overall concept is shown in Figure 2.

Complete isolation of different companies' data was ensured by MedChemica acting as a neutral third party and conducting all knowledge extraction through aggregation and filtering. This ensured that there could be no "drill down" from any rule to any companies' substructures or data by multiple layers of anonymisation.

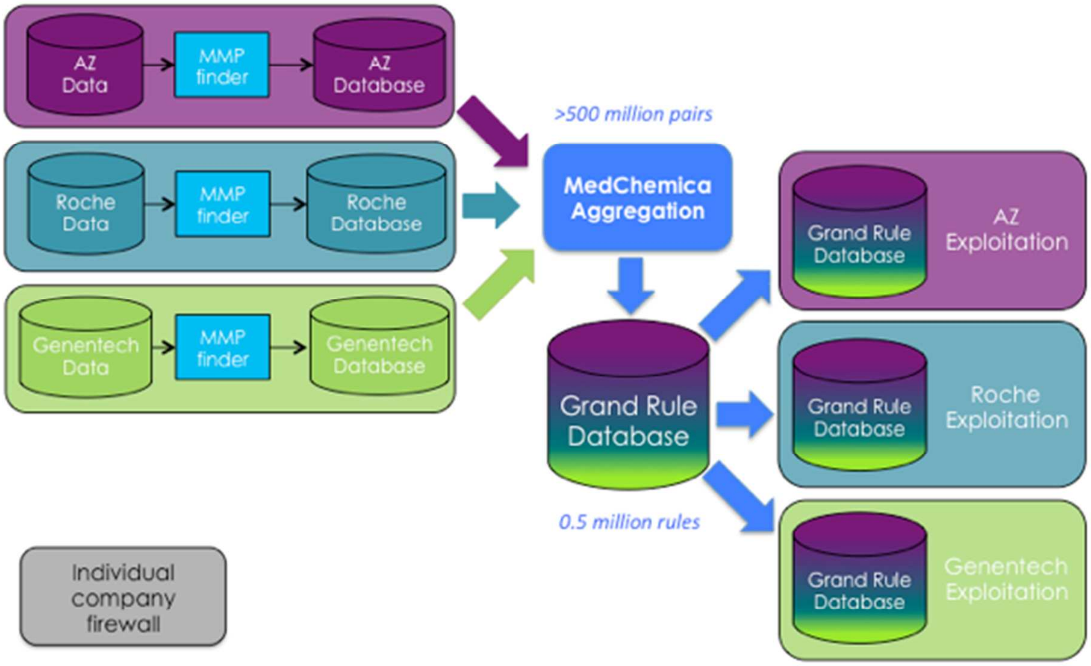


Figure 2: General assembly and use concept for the Grand Rule Database

To cover different possibilities, rule statistics were calculated in two different ways. Assuming that the differences between all pairs are normally distributed, median, standard deviation, and standard error of the mean was calculated for each transformation. Since many transformation involve censored data (*i.e.* measured values with ">" or "<" signs), non-parametric statistics based on a two-tailed binomial test were calculated to detect whether a given transformation significantly increases or decreases a property. Since the two-tailed binomial test requires at least six pairs for significance at the p=0.05 level, all rules with less than six pairs were discarded. An additional feature of this approach is that chemical transformations that have only been tried very few times will not be shared amongst contributing parties. There is a tension in this choice in that these may be "emerging" novel medicinal chemistry approaches which would be valuable to further explore, but in contrast these may be the “novel” medicinal chemistry approaches being most recently developed within a company and for which they wish to maintain an advantage. More details on the rule processing can be found in the supporting information.

When only one company contributed to an ADMET endpoint, the rules extracted by this process were added to the grand rule database (GRD). Where biological or physical chemistry endpoints could be compared between companies, the statistically significant transformations (according to the two-tailed binomial test) were extracted for calibration from each company set, and then the intersection set of transformations (those with identical canonical SMIRKS) was found. This intersection set was used to define calibration factors between the assays using a two-way Analysis of Variance (ANOVA) procedure.[24] A minimum of 8 transformations in the intersection set was required for calibration. Each of the transformations required a minimum of 6 matched pairs. If there were more than 200 transformations in the intersection set, a random subset of 200 transformations was used. As all transformations are mirrored, (A→B and B→A directions) for any given pair of transformations, only the positive median direction transformation was used to avoid a significant overestimation of the correlation between contributors as obviously delta A→B = -1* delta B→A. The mean change for the set of transformations, the means by contributor and the grand mean coefficients were taken from the ANOVA coefficients. The calibration factors for each contributor were calculated as the ratio of the grand mean over contributor mean. Each contributors data set was then scaled by the calibration factor and re-aggregated to generate a ‘common calibrated value’ representing the best estimates for each transformation. Merging all pairs with adjusted differences into rules gave a significant synergy benefit, where for example a transformation that had only 3 examples from one company, 2 from another and 1 from a third company could now pass the binomial test if all the examples changed the property in the same direction. The scaling procedure is schematically shown in figure 3.

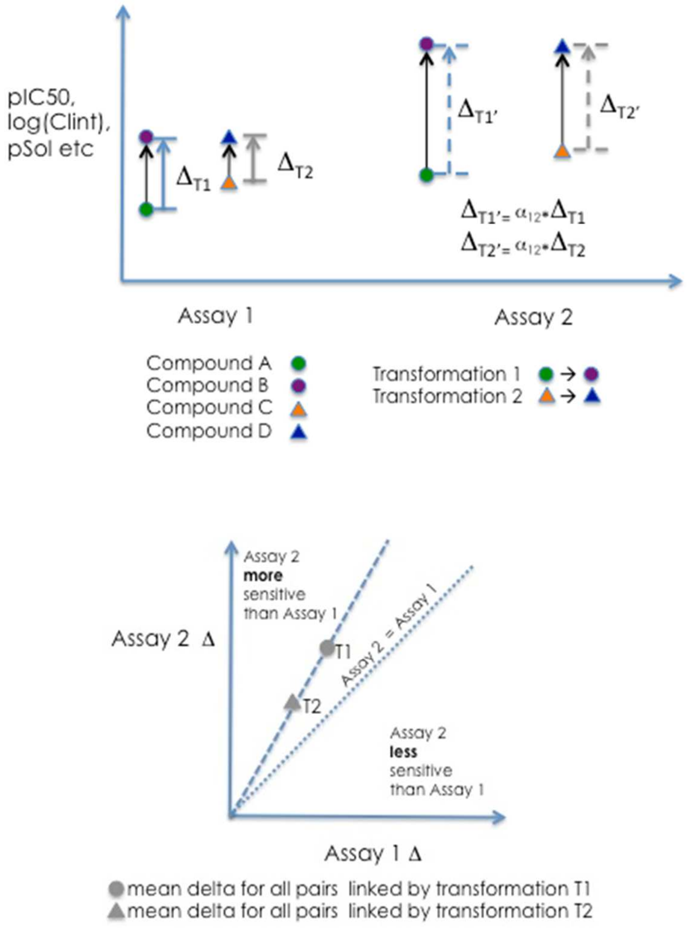

pIC50,
log(Clint),
pSol etc
Δ_{T1}
Δ_{T2}
$\Delta_{T1'}$
$\Delta_{T2'}$
$\Delta_{T1'} = \alpha_{12} \cdot \Delta_{T1}$
$\Delta_{T2'} = \alpha_{12} \cdot \Delta_{T2}$
Assay 1
Assay 2
Compound A ●
Compound B ●
Compound C ▲
Compound D ▲
Transformation 1 ● → ●
Transformation 2 ▲ → ▲
Assay 2 Δ
Assay 2 more sensitive than Assay 1
T1
T2
Assay 2 = Assay 1
Assay 2 less sensitive than Assay 1
Assay 1 Δ
● mean delta for all pairs linked by transformation T1
▲ mean delta for all pairs linked by transformation T2

Figure 3: Linearly displaced assays of different sensitivity

The statistical methods appropriate for this comparison have been developed in the clinical chemistry arena to allow different laboratories to compare their results against each other[25] and more generally in the meta-analysis of clinical trials.[26] This is equivalent to the established medicinal chemistry practice of recording the ranking of changes in extracting knowledge from publications or presentations; the exact values might be different but the ranking should be the same. It is reasonable to assume that scientists constructing assays in different organizations will have developed them such that they give the same ranking of standard compounds if not an identical numeric result.

Results

Rule gain

Over all the sets of assay data analyzed, 43 million unique structural transformations were found. 372,419 of these had more than six examples in any assay and passed the binomial test, each forming a medicinal chemistry rule. The distribution and overlap in rules per company is shown in figure 4.

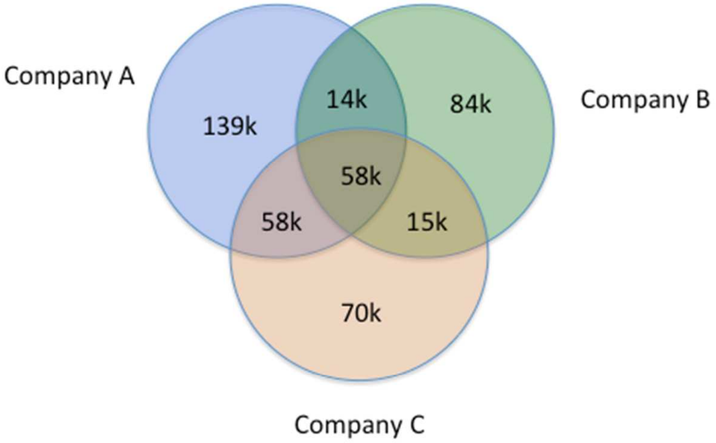


Figure 4: The origin of rules by company. The overlaps indicate the rules that contain examples from multiple companies, *i.e.* 58,000 rules had examples from all three companies, 139,000 rules were derived from company A data only.

A key question is: "what does a company gain from sharing data with the others". In numeric terms the average gain in rules for a company is 350% - *i.e.* if a company found 100 rules from analyzing its own data sets, it would have 450 rules by joining the consortium. On average this indicates how much non-overlap there has been between medicinal chemistry practice, with a large number of rules resulting from swapping knowledge from different areas. Also we can calculate the synergistic gain as the increase in number of rules found by merging the pairs and analyze the combined set minus the sum of what companies would have found individually, the average synergistic gain was 17%. This is the proportion of rules that had too little statistical support in any individual company, but “appear” by merging data. Some insight into the question “how large is the medicinal chemistry tool kit“ can also be addressed using this dataset, since the number of unique transformations can be extracted. For the 372,419 transformations that have been tried enough times to be assessed, there were 126,064 different A→B modifications when ignoring the local chemical environment. This also suggests that medicinal chemists in different companies are working in different areas of chemical space.[27–29] Analysis of the frequency of matched pairs showed a Zipfian distribution as previously reported by Hussain and Rea.[17] An example is shown for one company in the consortium’s contribution of human microsomal clearance data, of the 400,191 observed matched pairs, there were 303,966 unique SMIRKS. A breakdown of the numbers of examples of SMIRKS is shown in Figure 5. 86% of the transformations were only observed once, with only 1% being observed six or more times.

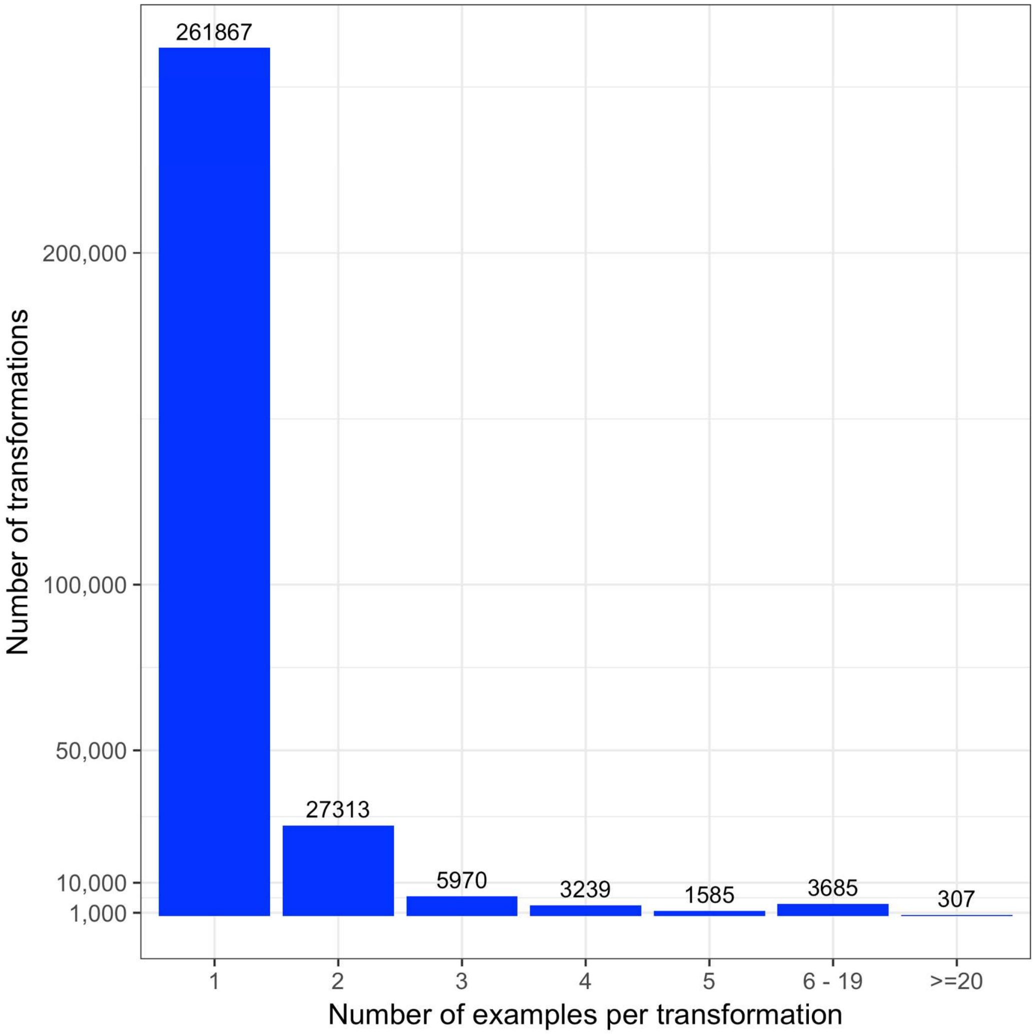


Figure 5: Distribution of number of pairs per transformation for a single company's human microsomal clearance data

Scaling factors

The distribution of calibration factors from all assay comparisons across all properties is shown in Figure 6. The calibration factors showed a median of 1.02; 80% of the calibration factors lie between 0.78 and 1.15 with two of the low outliers being assays where different technologies were used for signal detection. This suggests that, with appropriate care, the combination of data from multiple sources is feasible and that for most properties, different companies operating the same methods for a property determination generate comparable changes in measured values.

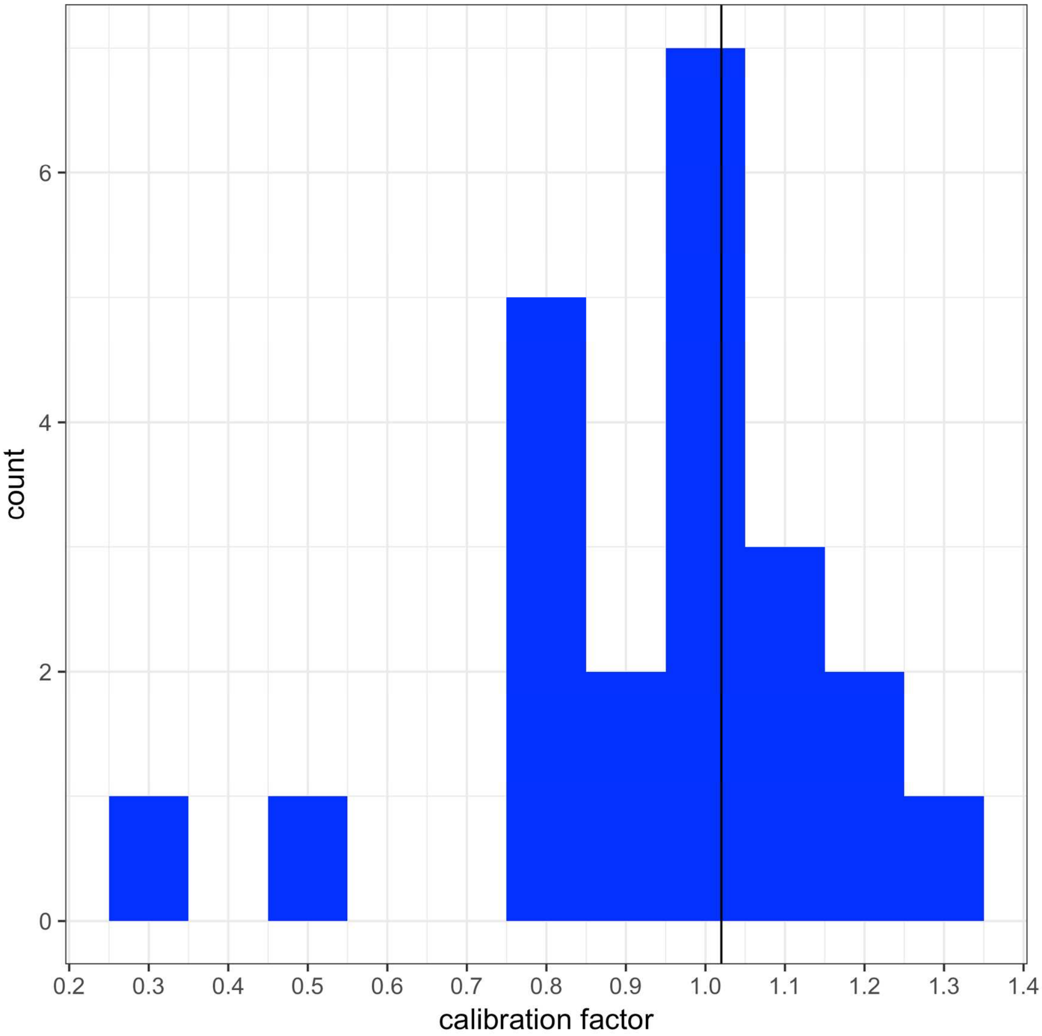


Figure 6: Distribution of the calibration statistics (median shown in black), count indicating the number of assays for a given bin.

Roche rules vs full dataset

One of the key questions for merging data between companies is how well these data agree with one another, or in other words, if and how much noise is added by merging the data. In our approach, we only formed pairs between compounds that have been measured in the same assay. It has been demonstrated that forming pairs between compounds that have been measured in assays that formally measure the same endpoint but are run by different operators in different places could increase the uncertainty of the rules dramatically, and the number of additional pairs gained is unlikely to compensate for the noise introduced.[15] For GRD, we compared the rules gained from a single company database (Roche as an example) versus the GRD rules (see figure 7) to compare the rules before and after the merging.

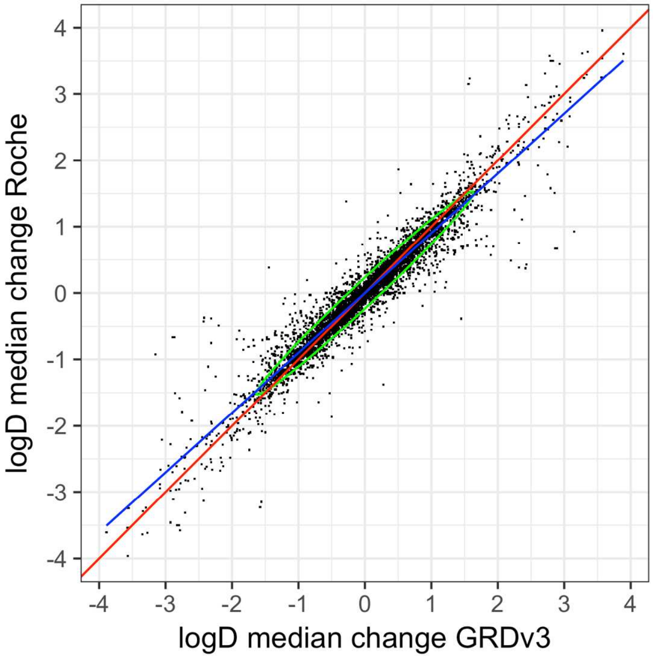


Figure 7: Correlation between predicted changes in logD based on Roche only rules and joint companies' rules. Overall R^2 is 0.98. Unless stated otherwise, in this and all succeeding scatterplots the pale green density ellipse contains 99%. The red line is the line of slope 1, intercept 0. The blue line indicates the linear fit. In this case the vast majority of the data lies on the 1:1 line between median logD change of -2 and +2 making the density ellipses hard to see.

For logD, the rules derived from Roche only data and the rules derived from the joint set of pairs compare extremely well with an R^2 of 0.98 and an RMSE 0.09. Overall, there are 255k logD rules in the GRD and 81,617 rules based on Roche data alone.

LogD is a relatively simple property to measure, depending on the equilibrium of a compound between two different solvents. Other important properties are more complicated, and their reproducibility and transferability between companies is likely lower. In Figure 8 we show the correlation between rules for human microsomal clearance derived from Roche data alone and the joint companies' data.

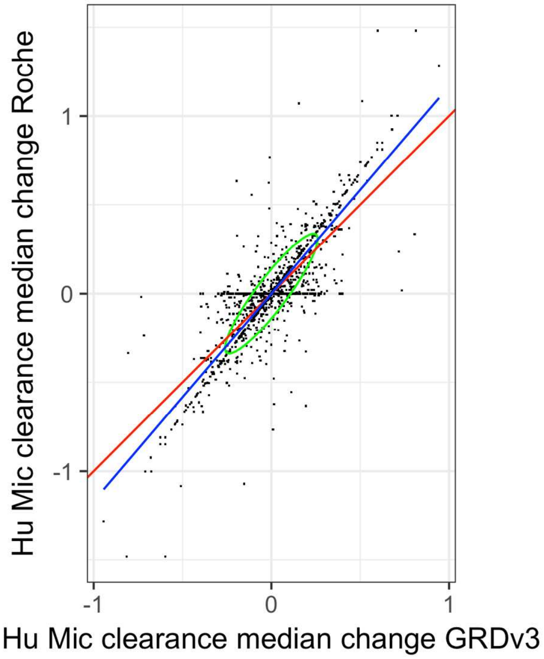


Figure 8: Comparison between Roche-only and GRD rules for human microsomal clearance. Overall R^2 is 0.76 and RMSE 0.11.

Overall, with $R^2 = 0.76$, the correlation for the microsomal clearance rules is lower than for logD. Nevertheless, almost all rules agree qualitatively, and only in 64 out of 933 the prediction changes from a negative to a positive median difference or vice versa (see Supporting Information on how positive, negative, and neutral effects are statistically defined in this work). Some of the rules in GRD shown in Figure 8 are based almost exclusively on Roche data. Those obviously lie close to the line of unity. For other rules, there are very few pairs from Roche only, and for a subset of those rules, the Roche contributing pairs had almost no change in clearance. The rules in that subset are visible as a horizontal line. Being more restrictive about the inclusion criteria for the comparison, those apparent lines disappear, but the overall correlation is the same (data not shown).

GRD rules vs previous rules

Since the overall comparison is very favorable, we were interested in finding out how individual rules from GRD compare to previously published rules. In 2015, Huchet *et. al.* published a paper on fluorine effects and their impact on physicochemical properties.[30] From prototypical examples given for logD, we extracted MMPs and compare them with the statistics for the closest transformation we could find in GRD (see table 1).

Fluorine can have a large effect on the pKa of a nitrogen in close vicinity. Therefore, transformations were selected with a balance between the number of pairs, chemical similarity, and being sure that no ionizable center is affected. We did not find contextually very close analogs of the compounds presented by Huchet *et. al.*. In the congeneric transformation series we found, an aromatic ethoxy group is increasingly fluorinated, whereas in the Huchet *et. al.* paper an aromatic propyl group is increasingly fluorinated. Nevertheless, the trend reported by Huchet *et. al.* is the same as the trend we find in GRD: a single fluorination reduces logD most strongly, whereas further fluorinations will reduce this effect.

Source	Transformation	ΔlogD ± std (nPairs)

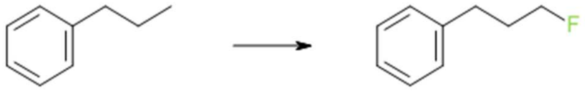
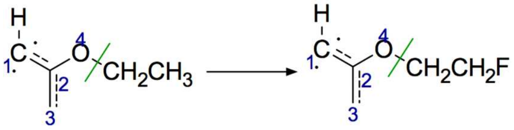
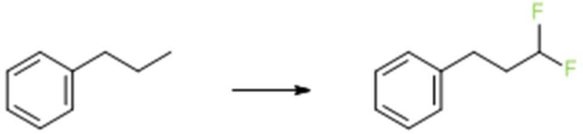
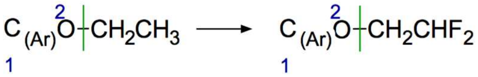
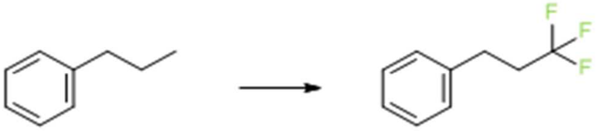
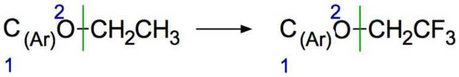
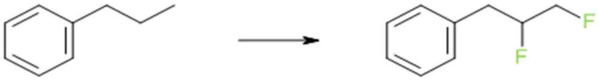
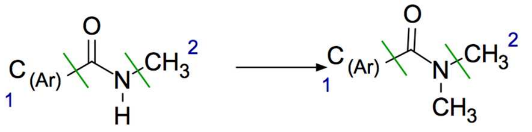
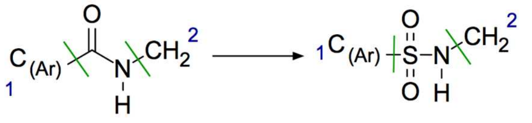
Huchet		-0.7
GRD	 [c:1]([H])[c:2]([c:3])[O:4]C([H])([H])C([H])([H])([H])>>[c:1]([H])[c:2]([c:3])[O:4]C([H])([H])C([H])([H])F	-0.48 ± 0.72 (8)
Huchet		-0.6
GRD	 [c:1][O:2][C]([H])([H])[C]([H])([H])([H])>>[c:1][O:2][C]([H])([H])[C]([H])([F])[F]	-0.11 ± 0.28 (18)
Huchet		-0.4
GRD	 [c:1][O:2][C]([H])([H])[C]([H])([H])([H])>>[c:1][O:2][C]([H])([H])[C]([F])([F])[F]	0.21 ± 0.59 (50)
Huchet		-0.8 - -1.0
GRD	insufficient examples	Insufficient examples

Table 1: Comparison between Fluorine effects published by Huchet *et. al.* and GRD statistics.

Here and in all following transformation depictions, the green bar indicates the bonds that separate the constant and the variable part of the transformation. The atoms that are depicted and belong to the constant part are the atoms defining the environment of the transformation.

As a second comparison, we looked at amide/ sulfonamide methylation. In one of the first MMP analysis papers, Leach *et. al.* statistically showed that amide N-methylation increases solubility.[2] Subsequently Ritchie *et. al.* presented statistics showing that on average, amide methylation on aromatic amides reduces logD and increases solubility.[31] They rationalize this behavior with the planarity-breaking effect of aromatic amide N-methylation. In contrast to amide N-methylation, they show that sulfonamide methylation always increases logD and reduces solubility. The GRD data is in perfect qualitative agreement with their observation (see table 2): In our data, aromatic amide N-methylation on average decreases logD by 0.25 log units and improves solubility by 0.26 log units. In contrast, aromatic sulfonamide N-methylation increases logD by 0.37 units and decreases solubility by 0.10 log units. Note that the rules used here are not merely hydrogen to methyl transformations, but rather specific linker transformations as depicted in table 2.

Transformation	ΔlogD ± std (nPairs)	ΔSolubility ± std (nPairs)
 O C (Ar) 1 N H CH3 2 O C (Ar) 1 N CH3 CH3 2 [c:1][C](=[O])[N]([H])[C:2]([H])([H])([H])>>[c:1][C](=[O])[N]([C:2]([H])([H])([H]))[C]([H])([H])([H])	-0.25 ± 0.45 (190)	0.26 ± 0.87 (144)
 O C (Ar) 1 N H CH2 2 1 C (Ar) O S O N H CH2 2 [c:1][C](=[O])[N]([H])[C:2]([H])([H])([H])>>[c:1][S](=[O])(=[O])[N]([H])[C	-0.19 ± 0.41 (64)	-0.10 ± 0.48 (23)

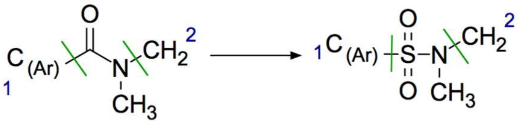
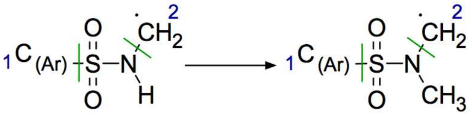
:2]([H])([H])		
 [c:1][C](=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])>>[c:1][S](=[O])(=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])	0.79 ± 0.21 (6)	insufficient examples
 [c:1][S](=[O])(=[O])[N]([H])[C:2]([H])([H])>>[c:1][S](=[O])(=[O])[N]([C]([H])([H])([H]))[C:2]([H])([H])	0.37 ± 0.35 (18)	-0.10 ± 0.67 (17)

Table 2: Effect of changing a secondary into a tertiary aromatic amide/ sulfonamide on solubility.

Another comparison with rules previously identified by Papadatos *et. al.*[9] is detailed in the supporting information Table S17. We generally find that the rule statistics found in GRD compare very well to previously published rule statistics. Differences result from different encoding of the environment, sampling (both number of pairs and chemical diversity), and inter-company variations in measuring different endpoints (see paragraph on Roche vs. joint rules).

Solubility vs logD

The most basic physicochemical parameters of relevance for drug design are solubility and logD. Initially we explored separating the thermodynamic and kinetic solubility assays. However we discovered that the assays, although known to give different absolute values, give the same compound rankings therefore rules could be inferred from combining the mixed solubility data sets. This is a quantitative equivalent to medicinal chemistry practice of inferring what are good “solubilizing groups” irrespective of solubility assay. LogD and solubility are often highly correlated, with the only other dominant factor apart from lipophilicity that influences solubility being the stability of the crystal structure, which can be measured by the melting point. [32] A plot of the correlation between the solubility and logD rules is shown in figure 9.

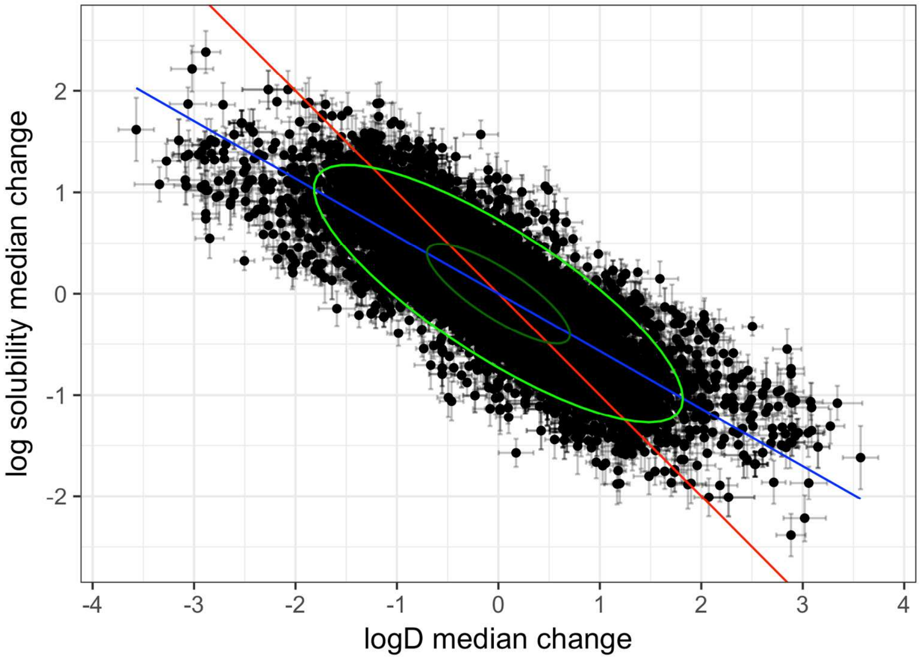


Figure 9: Solubility vs logD effects, >=20 pairs per rule, n=13453. R^2 = 0.66, slope = -0.57, intercept = 0. Red line: line of slope -1, intercept 0. Unless stated otherwise, in this and all the succeeding scatterplots, the dark green ellipse contains 50% of the data.

Figure 9 shows that logD and solubility are also highly correlated in transformation space. The R^2 is 0.66, with a slope of -0.57. If only rules with at least 50 pairs are considered, the R^2 increases to 0.72 (graph not shown). This means that on average a change of 1 in logD translates into a solubility change less than one order of magnitude. However the breadth of the distribution should be noted. For isolipophilic transformations, *i.e.* those with a ΔlogD of 0, the change in solubility can range over +/- 1 log unit.

There are a number of rules that have good statistical support but break the logD/solubility correlation, three examples are shown in table 3.

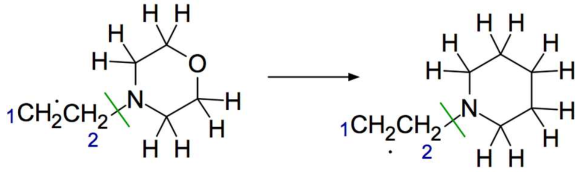
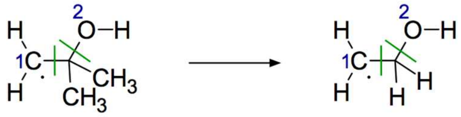
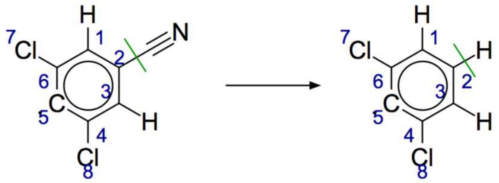
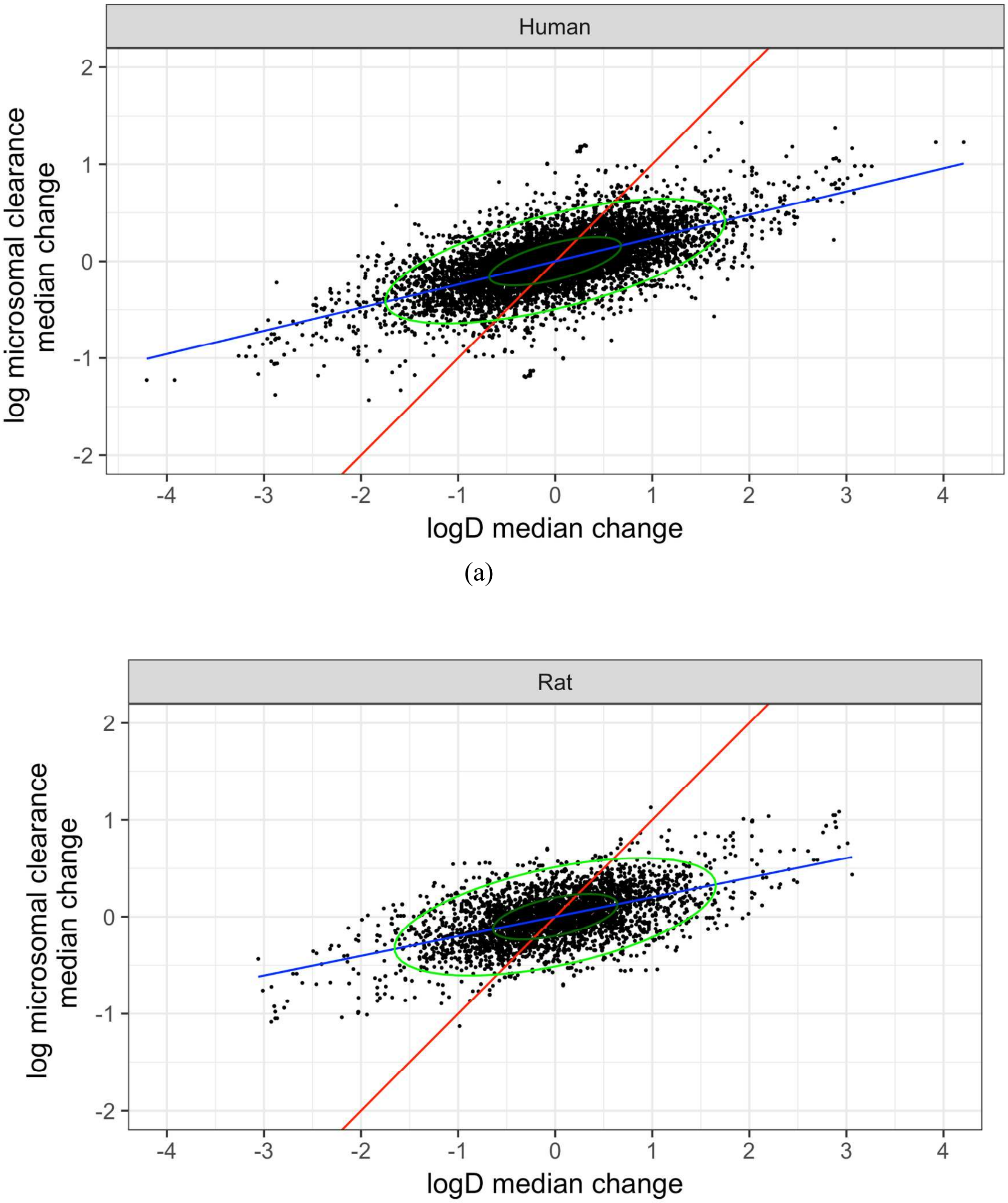
Transformation	ΔlogD ±std (nPairs)	ΔlogSol ±std (nPairs)
 [C:1]([H])([H])[C:2]([H])([H])[N]1[C]([H])([H])[C]([H])([H])[O][C]([H])([H])[C]([H])([H])1>>[C:1]([H])([H])[C:2]([H])([H])[N]1[C]([H])([H])[C]([H])([H])[C]([H])([H])[C]([H])([H])[C]([H])([H])1	0.00 ± 0.67 (91)	0.73 ± 0.72 (87)
 [C]([H])([H])([H])[C]([C]([H])([H])([H]))([C:1]([H])([H]))[O:2]([H])>>[C:1]([H])([H])[C]([H])([H])[O:2]([H])	-0.59 ± 0.49 (82)	0.03 ± 0.72 (98)
 7 1 2 3 6 5 4 8 7 1 2 3 6 5 4 8 [c:1]1([H])[c:2]([c:3]([H])[c:4]([c:5][c:6]1[Cl:7])[Cl:8])[C]#[N]>>[c:2]1([H])[c:1]([H])[c:6]([c:5][c:4]([c:3]1([H]))[Cl:8])[Cl:7]	0.45 ± 0.64 (50)	0.46 ± 1.02 (65)

Table 3: Example transformations where the change in logD and Solubility is not as one would expect from the logD-Solubility correlation.

The medians for the examples shown in Table 3 are statistically all very well backed, although there is quite some variation for the individual pairs as can be seen from the standard deviations. In the first example, an aliphatically N-connected morpholine is exchanged by an aliphatically N-connected piperidine.. In the second example, a tertiary dimethyl alcohol is transformed into a primary alcohol. This on average reduces logD by 0.59 log units, but has almost no effect on solubility. In the last example, a 4-cyano-2,5-dichlorophenyl is transformed into a 2,5-dichlorophenyl. This transformation increases logD by 0.45 log units, but it also increases solubility by 0.46 log units. From this and other examples (see supporting information) we have seen, it appears that aromatic cyano groups do not help for solubility. In the supporting information, some additional transformations with unexpected relationships between the effect on logD and solubility are exemplified in Table S5-S7.

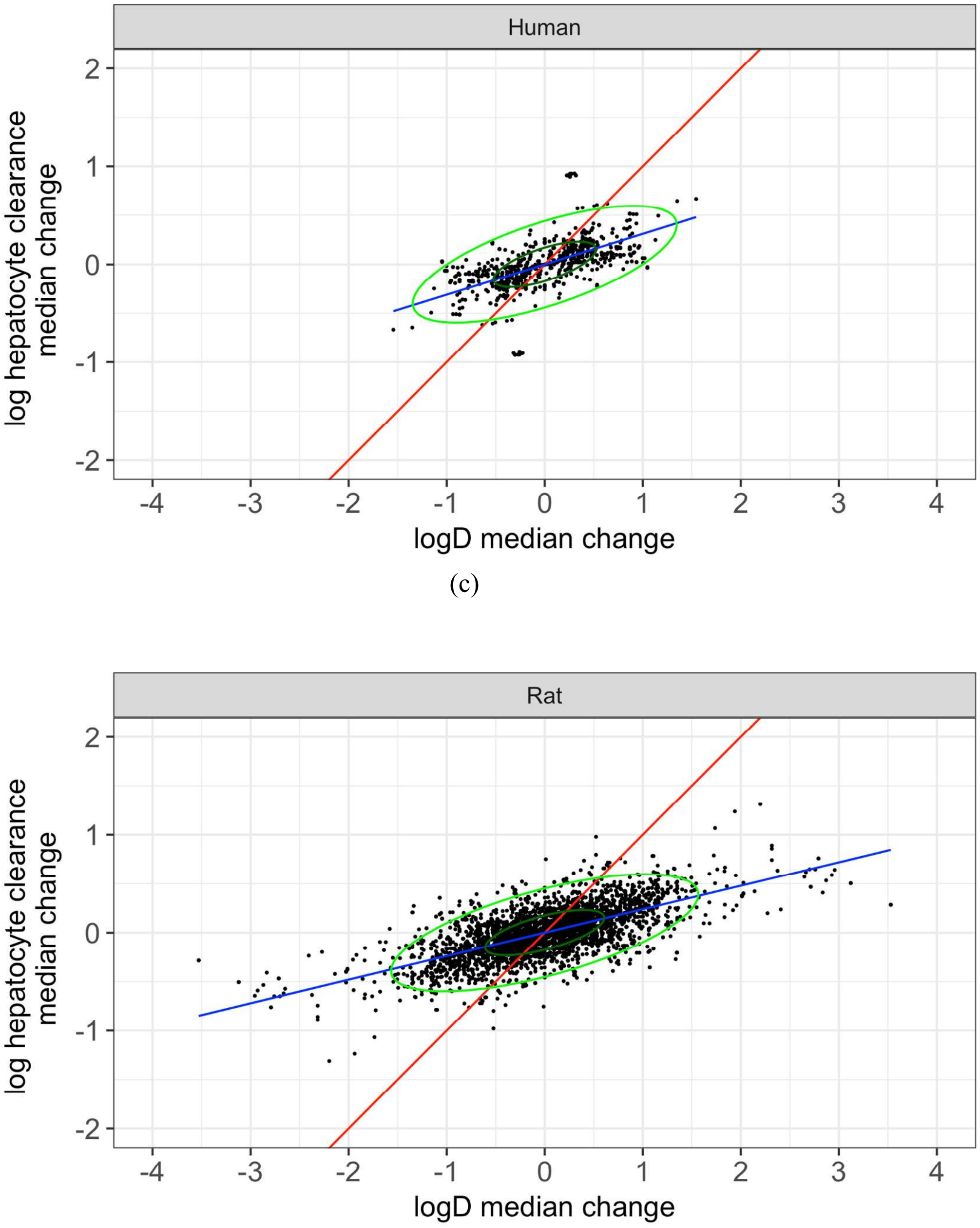
Clearance vs logD

Reducing lipophilicity is often considered as a common strategy to improve metabolic stability. In fact, 19 of GRD's top 20 rules (see Table S8 in supporting information) that decrease liver microsomal clearance and hepatocyte clearance are also accompanied by a logD decrease. Figure 10 shows a plot of the median change of microsomal and hepatocyte clearance in human and rat against the corresponding median change in logD where we have > 20 examples. (Microsomal and hepatocyte clearance data was used as pure Clint values, uncorrected for binding.) The plot shows a good correlation between these two properties (R^2=0.40 and 0.30 in human and rat microsomal clearance vs logD and 0.34 and 0.41 in human and rat hepatocyte clearance vs logD). The plot suggests that on average a change of 1 in logD translates into a human or rat in-vitro clearance change of 0.2 -0.4 log units.


Human
log microsomal clearance median change
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
logD median change
Rat
log microsomal clearance median change
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
logD median change

(a)

(b)



(c)

(d)

Figure 10: Median change of logD vs median change of (a) human (n=11,572, R^2=0.40, slope = 0.23) and (b) rat (n=5,056, R^2=0.30, slope = 0.20) liver microsomal clearance. Median change of logD vs median change of (c) human (n=812, R^2=0.33, slope = 0.31) and (d) rat (n = 4,937, R^2=0.41, slope = 0.24) hepatocyte clearance. Data is only shown for rules based on >=20 example pairs for every transformation.

There are some transformations that reduce human microsomal clearance while keeping logD neutral or even increasing it. Table 4 shows three such examples.

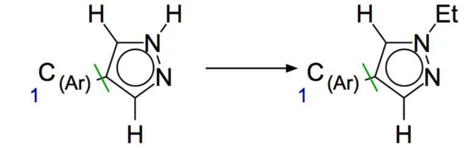
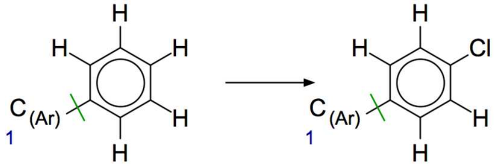
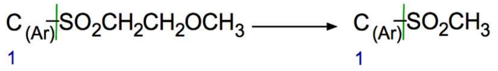
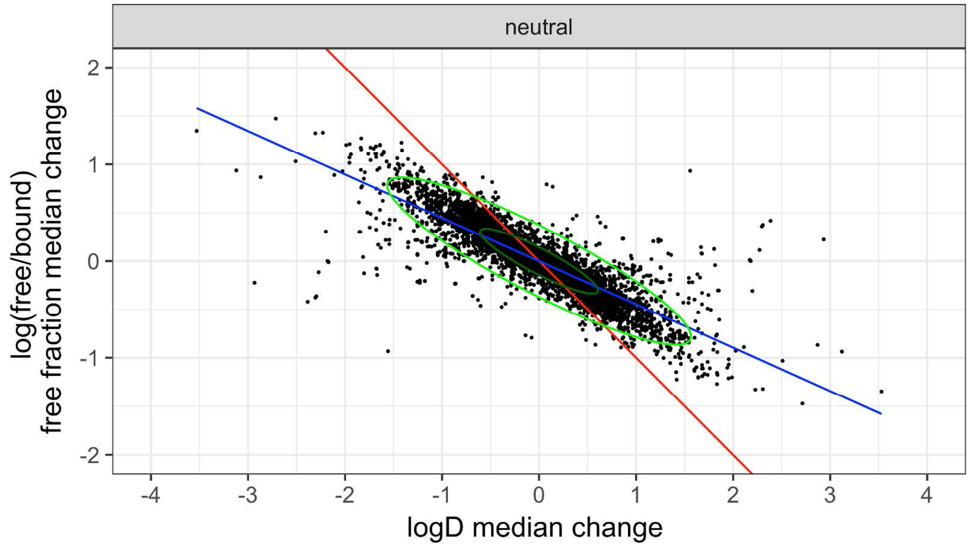
Transformation	Human microsomal Clearance median change ± std (nPairs)	logD median change ± std (nPairs)
 H H N C (Ar) 1 N H H Et N C (Ar) 1 N H [c:1][c]1[c]([H])[n]([H])[n][c]([H])1>> [c:1][c]1[c]([H])[n][n]([c]([H])1)[C]([H])([H])[C]([H])([H])([H])	-0.34±0.71 (13)	0.35±0.45 (15)
 H H H C (Ar) 1 H H H H Cl C (Ar) 1 H H [c:1][c]1[c]([H])[c]([H])[c]([H])[c]([H])[c]([H])1>> [c:1][c]1[c]([H])[c]([H])[c]([c]([H])[c]([H])1)[Cl]	-0.32±0.51 (53)	0.7±0.74 (117)
 [c:1][S](=[O])(=[O])[C]([H])([H])[C]([H])([H])[O][C]([H])([H])([H])>> [c:1][S](=[O])(=[O])[C]([H])([H])([H])	-0.59±0.38 (14)	0.0±0.11 (19)

Table 4: Transformations that decrease human microsomal clearance while keeping logD constant or even increasing it.

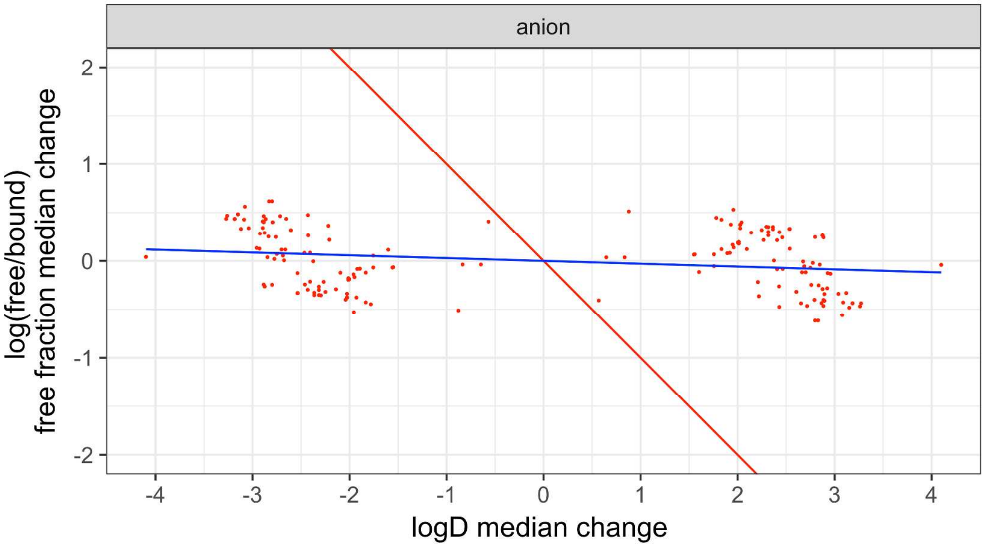
In the first transformation shown, a potentially reactive pyrazole-NH is protected by an ethyl. This increases logD by 0.34 log units, but at the same time on average reduces clearance by 0.34 log units.[33] In the second example, a terminal unsubstituted phenyl ring is substituted with chlorine in the para position. This leads to a drastic increase in logD by 0.7 log units, but also to an average decrease in clearance by 0.32 log units. Fluorine in the para position of a phenyl ring did not produce same statistical effect as chlorine; details of chlorine and fluorine substitution on a phenyl ring are in Table S9 in the supporting information. In the last transformation, an aromatic methoxy-ethyl-sulfone is replaced by a methyl sulfone. This on average does not have any effect on logD, but it strongly reduces clearance by 0.59 log units. In the supporting information (Table S10-12), more examples of transformations that reduce clearance while keeping logD constant or even increasing it are provided.

PPB versus logD

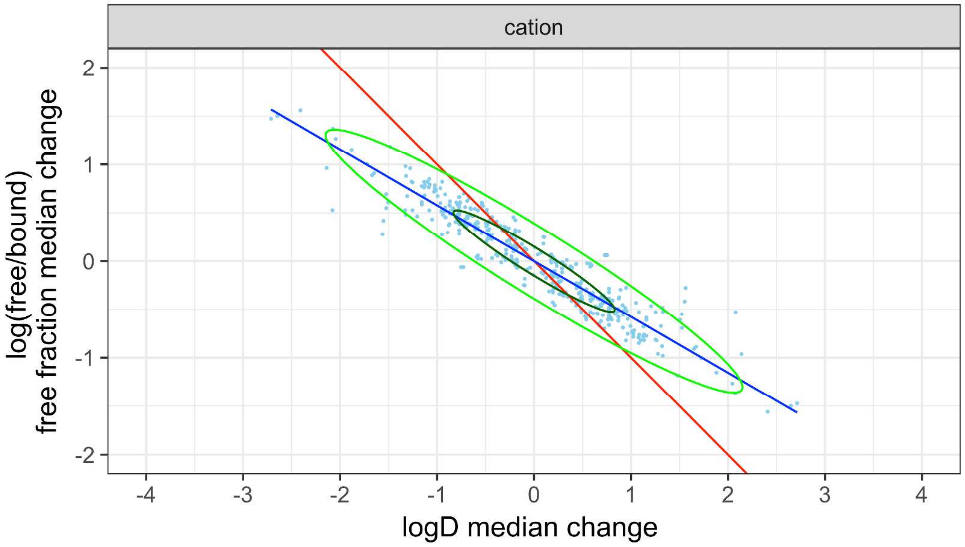
The optimization of free fraction as a goal is open to debate,[34] however when chosen as a strategy, reducing logD to increase the free fraction is an accepted approach.[35] We expect to observe the general trend between logD and free fraction in our data. In the GRD, changes in logD and human free fraction correlate fairly well for rules with more than 20 examples. Many outliers belong to the group of transformations where there is a change in anion count (colored blue in figure 11). Those are mainly neutral to acidic or acidic to neutral transformation changes, which have previously been reported to be “special” for free fraction.[36]


neutral
log(free/bound)
free fraction median change
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
logD median change

(a)


anion
log(free/bound)
free fraction median change
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
logD median change

(b)



(c)

Figure 11: human free fraction vs logD effect: red line = slope -1, intercept 0. (a) Greater than or equal to 20 pairs per rule and neutral: R^2= 0.71, Slope= -0.45, 7942 rules. (b) Rules with anion increase not equal to zero: R^2= 0.06, Slope= -0.03, 192 rules. (c) Rules with cation increase not equal to zero: R^2= 0.90, Slope= -0.58, 738 rules.

Figure 11 shows that differences in free fraction are highly correlated with differences in logD (R^2= 0.73, Slope= -0.46, 8,680 rules, not visible from these plots), if no anions (*e.g.* acids) are introduced or removed. This is a valuable finding, since it means that ΔlogD can be used as a good surrogate for human Δlog(free/bound), if no exchange of anions is involved. Despite have a huge effect on logD, the addition or removal of anions on average has no effect on PPB, as can be seen from figure 11b. The change in logD due to introduction or removal of cations, in contrast, is again highly correlated to the change in PPB, as can be seen from figure 11c. If the free fraction is low, it becomes very hard to measure, and improvements in free fraction can in many cases not be monitored experimentally since they are still below the limit of detection.

Some free fraction transformations are counterintuitive because logD is either increased or constant, but the free fraction increases nevertheless, and an acid is not involved in the transformation. Table 5 shows three such examples.

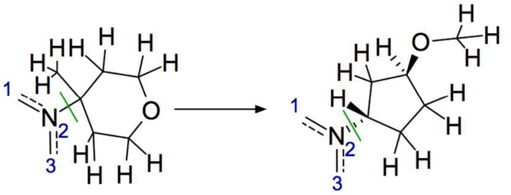
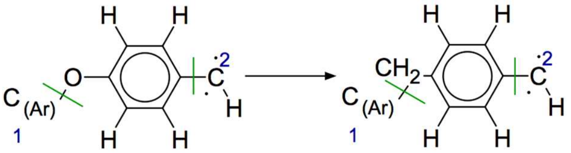
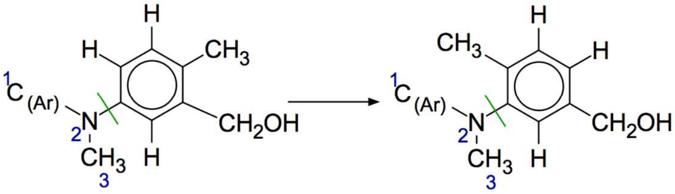
Transformation	ΔlogD ±std (nPairs)	HuΔlog (free/bound ±std (nPairs)
 [c:1][n:2]([c:3])[C]1([C]([H])([H])[C]([H])([H])[O][C]([H])([H])[C]([H])([H])1)[C]([H])([H])([H])>> [H][C@@]1([C]([H])([H])[C]([H])([H])[C@]([C]([H])([H])1)([H])[O][C]([H])([H])([H]))[n:2]([c:1])[c:3]	0.33 ± 0.14 (11)	0.32 ± 0.34 (11)
 [c:1][O][c]1[c]([H])[c]([H])[c]([c]([H])[c]([H])1)[C:2]([H])>> [c:1][C]([H])([H])[c]1[c]([H])[c]([H])[c]([c]([H])[c]([H])1)[C:2]([H])	0.48 ± 0.29 (11)	0.42 ± 0.28 (12)
 [c:1][N:2]([C:3]([H])([H])([H]))[c]1[c]([H])[c]([H])[c]([c]([c]([H])1)[C]([H])([H])[O]([H]))[C]([H])([H])([H])>>[c:1][N:2]([C:3]([H])([H])([H]))[c]1[c]([H])[c]([c]([H])[c]([H])[c]1[C]([H])([H])([H]))[C]([H])([H])[O]([H])	-0.1 ± 0.11 (21)	0.62 ± 0.19 (7)

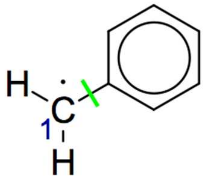
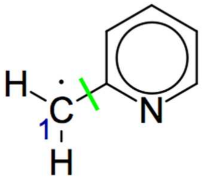
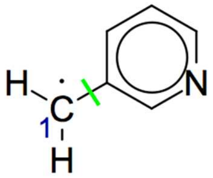
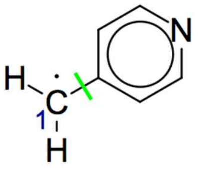
Table 5: Transformations that increase the free fraction and increase or keep logD almost constant.

The first transformation in table 5 shows the conversion of a 4-methyl tetrahydropyran into a methoxy-substituted cyclopentyl. This on average increases the free fraction by 0.32 log units, but also increases logD by 0.33 log units. The second transformation shows the conversion of an O-linker between two aromatic rings into a CH2-linker. LogD is increased by 0.48 log units, while the free fraction is increased by 0.42 log units. The last transformation in table 5 shows the shift of a methyl group on a substituted phenyl from the para to the ortho position. LogD is not significantly changed, but the free fraction is increased by 0.62 log units. More examples can be found in the supporting information Table S14-S16.

Functional group scorecards

Having a huge and diverse MMP database (over 350k unique structural transformations each with more than 6 example pairs) such as the GRD allows for a characterization of functional group replacements across many ADMET attributes. For frequently applied transformations such as the ones shown in Tables 6 and 7, almost complete "functional group scorecards" can be derived to predict the effect of a given frequent transformation on ADMET and PhysChem properties, normalized to a common starting structure. The scorecards give a quick overview of functional group properties across many different attributes relevant to drug design. Differences in the behavior of the pyridine and chloro-phenyl isomers is, as we might anticipate, smaller for the ADMET attributes where the interaction is less specific such as solubility, logD and protein binding, and larger for the more specific interactions such as P450 and hERG inhibition. Although we could speculate on the reasons for the regioisomer differences, detailed exploration of the causes of these differences is beyond the scope of this work.

Phenyl to pyridine transformations

			
Property	Δ ± std (nPairs)	Δ ± std (nPairs)	Δ ± std (nPairs)
logD*	-0.64 ± 0.70 (75)	-0.92 ± 0.72 (70)	-0.87 ± 0.79 (54)
Solubility	1.00 ± 0.75 (98)	0.84 ± 0.80 (98)	0.79 ± 0.81 (77)
hERG	-0.39 ± 0.47 (57)	-0.36 ± 0.44 (24)	0.12 ± 0.54 (37)
PPB human*	0.31 ± 0.42 (32)	0.73 ± 0.42 (37)	0.79 ± 0.38 (17)
human Mic Clearance	-0.11 ± 0.46 (43)*	-0.07 ± 0.57 (90)	Insufficient examples
human Hep Clearance	-0.28 ± 0.30 (8)	-0.14 ± 0.13 (5)	-0.04 ± 0.30 (7)
CYP3A4 inhibition	-0.31 ± 0.49 (22)	0.65 ± 0.54 (20)	0.96 ±0.67 (21)
CYP2D6 inhibition	-0.41 ± 0.51 (8)	0.49 ± 0.52 (7)	1.08 ± 0.66 (8)
CYP2C19	-0.69 ± 0.74 (3)	0.01 ± 1.11 (8)	0.74 ± 0.91 (6)

inhibition			

Table 6: Effects of changing a phenyl into a ortho-, meta-, and para -pyridine.* [R1][C] instead of [R1][C]([H])([H]) taken as environment due to paucity of pairs.

O- /m- /p- Chloro substitution on Phenyl

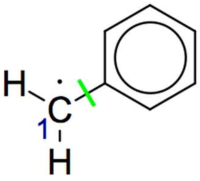
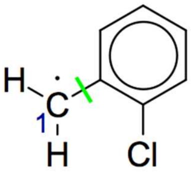
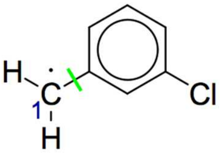
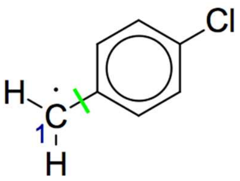
			
Property	Δ ± std (nPairs)	Δ ± std (nPairs)	Δ ± std (nPairs)
logD	0.48 ± 0.60 (66)	0.58 ± 0.57 (86)	0.67 ± 0.75 (127)
Solubility	-0.36 ± 0.77 (42)	-0.38 ± 0.75 (50)	-0.52 ± 0.85 (81)
hERG	-0.04 ± 0.29 (14)	0.16 ± 0.27 (5)	0.38 ± 0.35 (25)
PPB human	-0.33 ± 0.34 (32)	-0.49 ± 0.39 (51)	-0.54 ± 0.39 (68)
human Mic Clearance	0.24 ± 0.53 (47)	0.20 ± 0.47 (49)	0.00 ± 0.38 (61)
human Hep Clearance	0.16 ± 0.38 (7)	0.30 ± 0.29 (16)	-0.01 ± 0.37 (20)
CYP3A4 inhibition	0.29 ± 0.49 (20)	0.27 ± 0.46 (22)	0.27 ± 0.55 (31)
CYP2D6 inhibition	0.09 ± 0.36 (9)	0.20 ± 0.37 (12)	0.16 ± 0.54 (24)
CYP2C19 inhibition	0.32 ± 0.70 (5)	0.37 ± 0.61 (8)	0.16 ± 0.68 (14)

Table 7: Effects of ortho-, meta-, and para -chloro substitution on benzyls.

Discussion

We have determined whether medicinal chemistry ADMET knowledge can be shared and enhanced by combining transformations between organizations. Using the MMPA technology, knowledge about the effects on ADMET properties of medicinal chemistry transformations can be generated based on the joint pair datasets from all participating companies. The number of rules gained increases synergistically, since there are cases where the significance threshold to define a rule is only reached by adding all pairs linked by the same transformation from different companies.

Examining the effect of the transformations on human liver microsomal stability and LogD common to all contributing organizations to those from Roche alone shows excellent agreement. This shows that the merging data by the strategy presented in this contribution supplements inhouse rules much more than it increases the noise in the rule statistics due to mixing pairs measured at different companies. Although an exact quantitative comparison with previously published rules is not possible due to different MMPA definitions used in different publications, we observe qualitative agreement between the GRD rules and published rules. At first sight the synergistic gain may appear purely quantitative (more rules), it is important to realize, however, that the more pairs support those rules, the better they become. Therefore, ‘quantity has a quality all its own’ in MMPA. Also, if the pairs come from different companies, the structural variety in the underlying pairs most probably increases, making the rules more robust. We did not quantify this here, but we believe that in future work it would be valuable to develop a metric for quantifying the structural diversity in the underlying pairs for each rule as an indicator of robustness.

The correlation between logD and solubility in transformation space, *e.g.* the space that is relevant for medicinal chemistry optimization, has an R^2 of 0.66 and agrees very well with previously reported correlations for individual compounds. As such, this experiment confirms general basic medicinal chemistry principles. However, there are also a number of exceptions to the general logD-solubility rule. We here presented three statistically very well supported exceptions as examples, and many more can be found in the GRD. With MMPA-based rule databases such as the GRD, medicinal chemists can easily access and use all the “exception” rules in prospective design.

We also showed that clearance is correlated to logD in transformation space, although the correlation is much weaker ($R^2 = 0.3 – 0.5$). Clearance is a topic particularly well suited to MMPA, since beyond simple logD and exposure rules, principles to address clearance are almost exclusively based on rules about fragments and their relative stability. Here, we present only three rules that go counter to the accepted logD correlation. The supporting information contains some more rules, but the most natural way to access all the rules is through MMPA-based rule databases, since there are so many rules that it is impractical to print them all in a traditional article. With our big dataset, novel and interesting analyses about the correlations between species and microsomal and hepatocyte clearance can be made, all of which are however beyond the scope of this contribution.

As a third example, we present the analysis of the correlation between logD and PPB in transformation space. We find that the two are highly correlated, if changes in the number of anions (mostly acids) are discarded. This is an important finding: it means that in the cases where free fraction is very hard to measure, a first estimate can be made using a matched pair measurement of logD and the correlation between delta logD and delta log free fraction. MMPA was critical in establishing this relationship as normal plots of free fraction versus logD are often compromised by the noise generated by very high or low free fraction values and the limit of PPB detection. The real relationship between changes in logD and free fraction could only be revealed thanks to the size of the datasets and the number of rules in GRD.

Relationships between rules are also important and have been explored in the related matched molecular series studies.[38] With a growing body of well-defined rules across different properties available for the common transformations, it is possible to create functional group scorecards, such as shown above for phenyl Cl- and N-substitution (Tables 6 and 7). These scorecards enable a quick overview of functional group properties (relative to a joint standard) and inform the choice of functional group substitution for property optimization. Those can form the basis of analysis across the medicinal chemistry knowledge space that may allow induction of A→C rules from the known relationships of A→B and B→C. The recent work of Kramer[10] suggests that this may be possible though the contributing transformations need to be very well exemplified to avoid the amplification of experimental error when combining rules. Larger datasets than the one considered here may be needed to enable these enhancements.

Finally, such a corpus of knowledge may be used for the training of medicinal chemists. Historically medicinal chemists have learnt though their work on projects predominantly experientially, a time consuming and potentially unreliable route to gaining expertise.[39] With the possibility of encoding medicinal chemistry knowledge in a consistent, robust, data dense, sharable manner, there is an opportunity to accelerate and enhance medicinal chemists' skills. Access to a reference guide to what outcomes are "reasonable" for a given chemical change can allow chemists to set proposed compounds within the context of precedents and rank compounds in both synthetic tractability and probability of making the desired ADMET change.

Conclusions & Outlook

We have demonstrated that significant knowledge can be extracted from large scale, unsupervised mining of in vitro ADMET data from multiple pharmaceutical companies. Within the organizations collaborating in this work and in those who have been able to exploit the data in trial form in universities and not for profit organizations, the knowledge has contributed to solving a number of drug hunting project problems leading to several publications .[40–42] We have also shown examples of transformations that run counter to established wisdom. These may be useful tools to extricate a drug hunting team from a tight optimization corner and more generally to provide opportunities to enhance our understanding of the general SAR of the system being studied. Both of these are highly desirable goals.

It is conceivable that the ensemble of rules found through MMPA on large datasets (as large as the one studied here or even larger) could form the basis for a statistics-based reference for medicinal chemistry, an "encyclopedia of medicinal chemistry tactics".[37] In this case, it matters less that the underlying pairs have not all been measured at the same place. What is really important is to maximize the size of dataset to broaden the resulting corpus of rules and the statistical meaningfulness. For PPB, we obtained promising results in an analysis based on the partitioning by environment and ion class. Further partitions by factors such as shape or electronic descriptors could also be considered, though a clear mechanistic rationale justifying every partition is necessary as the analysis could fall into the traps of "data dredging" and generate false rules from statistical artifacts.

Peter Norvig, a director of research at Google has published on the "Unreasonable Effectiveness of Data",[43] the concept that with a very large quantity of data, useful knowledge can be extracted in the absence of an underlying mechanistic model. For simple in vitro ADMET properties it seems reasonable to suggest that we may be approaching this point in medicinal chemistry. It is possible to view the “brute force engineering” approach of data mining as inelegant to those more persuaded by a reductionist-theoretical approach to compound optimization, however we see these two approaches as being on a continuum. Established theories are often built on a modest amount of experimental data, whereas data mining transformations brings a vast depth of data with no mechanistic model beyond the “fundamental belief” of chemistry that “structure defines properties”. The opportunity to enhance our SAR theories for ADMET properties by using the learning from data mining is an area as yet under-developed and has the possibility of further accelerating drug hunting. As researchers at Pfizer have described it, this is the process of generating tacit knowledge from large datasets.[44] Both the reductionist-theoretical and pragmatic data mining approaches should accelerate lead optimization. Given that the urgent and essential goal of drug discovery is to deliver safe effective compounds to the clinic, and the continuing decrease in drug hunting productivity[45] the simple truth is that medicinal chemists must increasingly look to more efficient practices and reliance on robust predictions and substantiated rules.

Supporting Information

The supporting information contains the following information:

- Additional detail and diagrams showing the capture of the local chemical environment round a point of change in a transformation
- The data merging approach
- Flowchart showing the method for assigning rules as significant
- Solubility vs logD plots and confusion tables
- 72 additional rules in SMIRKS and drawn transformations

- Comparison to 7 previously published rules
- Single company vs GRD comparison confusion tables.

Corresponding Author Information

AstraZeneca:	Attilla.Ting@AstraZeneca.com	+44 1625 234706
Genentech	zheng.hao@gene.com	+1 650 467 7446
MedChemica	ed.griffen@medchemica.com	+44 1625 238843
Roche	christian.kramer@roche.com	+41 61 6822471

Author Contributions:

The key authors contributed the work equally.

Abbreviations Used:

GRD	Grand Rule Database
hERG	human ether-a-go-go related gene product
MMP	Matched Molecular Pair
MMPA	Matched Molecular Pair Analysis
MDCK cells	Madin Darby Canine Kidney cells

Acknowledgments

Sharing large quantities of knowledge across multiple large organizations is a rare occurrence in the pharmaceutical industry and would only have been possible by the support of a large technical team behind the scientists directly involved. At Genentech, Ying Littman and Slaton Lipscomb provided vital technical support and Ignacio Aliagas, Alberto Gobbi, Fabio Broccatelli, Jason Burch are thanked for many good discussions. We are also thankful to Steve St-Gallay and John Cumming for their assistance during the early stages of the collaboration. The support of the Openeye technical staff and Anthony Nicholls was highly valuable in delivering the cheminformatics aspects of the collaboration. The scale of the knowledge sharing undertaken would not have been possible without the support of key leaders within the businesses concerned and their role in enabling such an undertaking is gratefully acknowledged. In particular MedChemica is grateful to Andy Barker and Phil Jewsbury for support at the genesis of this work, and Mike Snowden, Susan Galbraith, Mene Pangalos and Luca Santarelli for helping prosecute the first phase of data sharing.

Biographies

Christian Kramer (ORCID-ID: 0000-0001-8663-5266)

Christian received his PhD from the University of Erlangen, Germany, working on the machine-learning prediction of physicochemical and biological properties of small molecules. In 2010, he moved to Basel, Switzerland, as a Novartis Presidential PostDoc, and in 2013 he accepted an assistant professorship position for Theoretical/ Computational chemistry at the University of Innsbruck, Austria. In 2015, he moved to F. Hoffmann-La Roche Ltd., where he works on therapeutic projects in multiple disease areas and develops computational methods. Christian's research interests lie in the field of computer-aided drug design, trying to map and support the complex rationalization and decision processes that underlie medicinal chemistry. This includes techniques such as QSAR, QSPR, matched-molecular pair analysis, data mining, and the analysis of experimental uncertainty.

Attilla Ting (ORCID-ID: 0000-0002-1590-1165)

Attilla Ting is a Computational Chemist and a project manager working with the Oncology Innovative Medicines drug discovery unit at AstraZeneca. Attilla's responsibilities focus around the molecular modeling support of a wide range of early and late stage projects along with project management responsibility. Attilla has a particular interest in structure-based drug design and automatic matched pair analysis. Attilla has published a number of papers in the drug discovery field. Attilla was awarded an Undergraduate degree in Chemistry and Computer science from The University of Leeds and continued into Postgraduate studies within the Chemistry Department to gain a PhD in Computational Chemistry with Dr Peter Johnson before joining AstraZeneca in 2000.

Hao Zheng (ORCID-ID: 0000-0002-1234-1801)

Hao Zheng is an Associate Scientist at Genentech, where he has worked for the last 11 years as a cheminformatics project manager and computational chemist in computational drug design group. His role involves developing software solution to streamline research workflow and provides computational modelling support to small molecule therapeutic projects. Hao has a particular interest in applying machine learning and predictive analytics in drug discovery. Hao holds a Masters degree in organic chemistry from Nanjing University and a Master degree in computer science from Illinois State University.

Jérôme Hert (ORCID-ID: 0000-0003-3062-8029)

Jérôme Hert received his PhD in cheminformatics from the University of Sheffield under the supervision of Peter Willett. He continued to develop chemogenomics approaches as a

Marie Curie Fellow in the laboratories of Brian Shoichet at UCSF, and of Didier Rognan at the University of Strasbourg. He joined F. Hoffmann-La Roche in 2009 and initially worked on therapeutic projects in multiple disease areas. In 2013, he was appointed Head of the Computer-Aided Drug Design section. A member of the Chemical Biology Leadership Team, he is responsible for the coordination of cheminformatics and molecular design contributions. Jérôme has experience and interest in the development and application of a variety of data mining, cheminformatics, and modeling approaches. His scientific contributions include over 30 scientific articles and patents.

Torsten Schindler

Torsten Schindler studied Chemistry at the Friedrich-Alexander University in Erlangen-Nuernberg, where he also got his PhD at the Computer Chemistry Center in the group of Prof. Timothy Clark. From 2000 to 2002 he worked as PostDoc in the field of Computational Chemistry at the Novartis Institute of Biomedical Research and got a permanent employee in 2002. In 2008 he moved to Basel and joined F. Hoffmann-La Roche as an Information Analyst in Pharma Research and Early Development Informatics and is currently working in the area of Data Science since 2015.

Martin Stahl

Martin Stahl studied chemistry in Freiburg and Würzburg, Germany, and obtained his PhD in theoretical chemistry in Marburg with Prof. Gernot Frenking. He joined Roche in 1997. Since then, he has held various leadership roles in computational and medicinal chemistry, biophysics and biostructure. Since 2015, he focuses on portfolio work and has led Program Management for Small Molecule Research at Roche pRED. He is an Advisory Board Member of Journal of Medicinal Chemistry, an editorial board member of

ChemMedChem and a Trustee of the CCDC. He is the recipient of the 2014 ACS National Award for Computers in Chemistry and Pharmaceutical Sciences

Graeme Robb (ORCID-ID: 0000-0002-4531-4375)

Graeme Robb is an Associate Principal Scientist at AstraZeneca, where he has worked for the last 15 years as a computational chemist. He currently works within the Innovative Medicines drug discovery unit, in the pursuit of drugs to fight cancer. His role involves using 3D molecular structure of drugs and their target proteins to simulate these systems and design better, drug-like molecules. Graeme has been a key player in establishing numerical and statistical learning methods for modelling and predicting the properties of molecules using their structural features. He holds a Ph.D. and a Masters degree in chemistry from the University of Edinburgh.

James J. Crawford (ORCID-ID: 0000-0002-6408-8246)

Dr. James Crawford is a Senior Scientist and Project Team Leader in Small Molecule Drug Discovery at Genentech. Born and raised in Glasgow, Scotland, James obtained his MSci and Ph.D. degrees in chemistry from the University of Strathclyde, while building an affinity for watching Rangers F.C. and listening to The Smiths. He joined Professor K.C. Nicolaou's laboratory at The Scripps Research Institute as a Fulbright scholar, then started his industrial career in 2006 with AstraZeneca in their Respiratory and Inflammation group. In 2010, he moved to Genentech, where he has worked across three disease areas – Immunology, Oncology and Antibacterials. One of his key roles at Genentech has been as the chemistry leader of the BTK project team at Genentech.

Jeff Blaney (ORCID-ID: 0000-0001-9505-552X)

Jeff received his Ph.D. in Pharmaceutical Chemistry from UCSF. He has worked in several Pharma, Biotech, and software companies in computer-assisted drug discovery and chemical informatics, focusing on structure-based design. He has lead Genentech's Computational Chemistry and Cheminformatics group in Small Molecule Discovery since October 2007. He has over 50 publications and patents.

Shane Montague (ORCID-ID: 0000-0003-3211-6422)

Shane Montague obtained his PhD from the University of Salford, Manchester. He has taught courses on information security and distributed systems at undergraduate and postgraduate level. He undertook postdoctoral research in European projects and has collaborated with international partners from across industry and academia in both the finance and energy sectors. In 2012, he joined MedChemica Limited as a computer scientist. He has implemented tools and systems that support the decision-making of medicinal chemists in drug discovery. His interests lie in the area of data-intensive applications, ranging from theory to design to implementation.

Andrew G Leach (ORCID-ID: 0000-0003-1325-8273)

Andrew Leach obtained his PhD from the University of Cambridge and undertook post-doctoral research at UCLA, supported by the Fulbright scheme. He returned to the UK to join AstraZeneca as a computational chemist and worked in oncology, diabetes and obesity. In 2012, he became one of the co-founders of MedChemica limited and also joined Liverpool John Moores University as a lecturer in the School of Pharmacy and Biomolecular Sciences. His interests include all aspects of computation to support drug discovery and medicinal chemistry as well as the use of quantum mechanics to unravel interesting reaction mechanisms. He has published more the 70 articles, 5 book chapters and is a named inventor on 12 patents.

Al Dossetter (ORCID-ID: 0000-0002-4181-3193)

Al Dossetter gained his PhD from Nottingham University and after post-doctoral research at Harvard University joined AstraZeneca. He has 13 years of experience in medicinal chemistry spread across oncology (hormonal and kinase inhibitors), inflammation (OA and RA, enzyme inhibitors and GPCR targets) and diabetes (obesity, GPCR and enzyme inhibitors). In 2012, he co-founded MedChemica Limited to use Matched Molecular Pair Analysis to accelerate medicinal chemistry. MedChemica now licenses a suite of software tools for companies to extract and share knowledge from their own data and combine with public data. The software and methodologies have been used by many pharmaceutical companies, universities and biotechs to accelerate drug discovery programmes. Al is an enthusiastic advocate for his science and is frequently invited to present his research.

Ed Griffen (ORCID-ID: 0000-0003-0859-554X)

Ed Griffen obtained his PhD from Imperial College, London and undertook post-doctoral research at the University of Waterloo-Kitchener, Canada. He joined Zeneca Pharmaceuticals in medicinal chemistry working in the CNS, infection, oncology and chemical biology areas. Taking a secondment into the computational chemistry group he co-developed matched molecular pair tools to quantify medicinal chemistry approaches. In 2012 he co-founded MedChemica Ltd, a company dedicated to improving medicinal chemistry practice. He has taught medicinal chemistry courses at the University of Manchester and AstraZeneca courses in the UK, France and India. He is a named inventor on 16 patents and co-authored more than 15 articles, book chapters and a textbook. His interests are in developing data driven methods to support decision making and medicinal chemistry education.

References

(1) Dossetter, A. G.; Griffen, E. J.; Leach, A. G. Matched Molecular Pair Analysis in Drug Discovery. *Drug Discovery Today* **2013**, *18* (15-16), 724–731.

(2) Leach, A. G. 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.

(3) Wassermann, A. M.; Bajorath, J. Large-Scale Exploration of Bioisosteric Replacements on the Basis of Matched Molecular Pairs. *Future Med. Chem.* **2011**, *3* (4), 425–436.

(4) Hajduk, P. J.; Sauer, D. R. Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *J. Med. Chem.* **2008**, *51* (3), 553–564.

(5) Wassermann, A. M.; Dimova, D.; Iyer, P.; Bajorath, J. Advances in Computational Medicinal Chemistry: Matched Molecular Pair Analysis: Matched Molecular Pair Analysis. *Drug Dev. Res.* **2012**, *73* (8), 518–527.

(6) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Methods and Principles in Medicinal Chemistry*; Oprea, T. I., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, FRG, 2005; pp 271–285.

(7) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool: Miniperspective. *J. Med. Chem.* **2011**, *54* (22), 7739–7750.

(8) Matlock, M.; Swamidass, S. J. Sharing Chemical Relationships Does Not Reveal Structures. *J. Chem. Inf. Model.* **2014**, *54* (1), 37–48.

(9) Papadatos, G. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity. *J. Chem. Inf. Model.* **2010**, *50*, 1872–1886.

(10) Kramer, C.; Fuchs, J. E.; Liedl, K. R. Strong Nonadditivity as a Key Structure–Activity Relationship Feature: Distinguishing Structural Changes from Assay Artifacts. *J. Chem. Inf. Model.* **2015**, *55* (3), 483–494.

(11) 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.

(12) Dossetter, A. G.; Douglas, A.; O'Donnell, C. A Matched Molecular Pair Analysis of in Vitro Human Microsomal Metabolic Stability Measurements for Heterocyclic Replacements of Di-Substituted Benzene Containing Compounds - Identification of Those Isosteres More Likely to Have Beneficial Effects. *Med. Chem. Commun.* **2012**, *3*, 1164–1169.

(13) 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* (16), 5906–5919.

(14) Dossetter, A. G. A Statistical Analysis of in Vitro Human Microsomal Metabolic Stability of Small Phenyl Group Substituents, Leading to Improved Design Sets for Parallel SAR Exploration of a Chemical Series. *Bioorg. Med. Chem.* **2010**, *18*, 4405–4414.

(15) Kramer, C.; Fuchs, J. E.; Whitebread, S.; Gedeck, P.; Liedl, K. R. Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. *J. Med. Chem.* **2014**, *57* (9), 3786–3802.

(16) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (1), 125–134.

(17) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50* (3), 339–348.

(18) Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. *J. Chem. Inf. Model.* **2010**, *50* (8), 1350–1357.

(19) Python Software Foundation. *Python Version 2.7.1*; Available at http://www.python.org.

(20) *OECHem Tookit*, 1.7.4.3 ed.; OpenEye Scientific Software, Santa Fe, NM. OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

(21) *MySQL 5.1.60, Client 14.14*; Oracle Corporation.

(22) R Core Team. *R: A Language and Environment for Statistical Computing, Version 3*; R Foundation for Statistical Computing: Vienna, Austria, 2015.

(23) *rpy2, R in Python*.

(24) Glantz, S. A.; Slinker, B. K. *Primer of Applied Regression & Analysis of Variance*, 2nd ed.; McGraw-Hill, Medical Pub. Division: New York, 2001.

(25) Rose, C. E.; Romero-Steiner, S.; Burton, R. L.; Carlone, G. M.; Goldblatt, D.; Nahm, M. H.; Ashton, L.; Haston, M.; Ekstrom, N.; Haikala, R.; Kayhty, H.; Henckaerts, I.; Durant, N.; Poolman, J. T.; Fernsten, P.; Yu, X.; Hu, B. T.; Jansen, K. U.; Blake, M.; Simonetti, E. R.; Hermans, P. W. M.; Plikaytis, B. D. Multilaboratory Comparison of Streptococcus Pneumoniae Opsonophagocytic Killing Assays and Their Level of Agreement for the Determination of Functional Antibody Activity in Human Reference Sera. *Clin. Vaccine Immunol.* **2011**, *18* (1), 135–142.

(26) *Introduction to Meta-Analysis*; Borenstein, M., Ed.; John Wiley & Sons: Chichester, U.K, 2009.

(27) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679.

(28) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 374–380.

(29) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50.

(30) Huchet, Q. A.; Kuhn, B.; Wagner, B.; Kratochwil, N. A.; Fischer, H.; Kansy, M.; Zimmerli, D.; Carreira, E. M.; Müller, K. Fluorination Patterning: A Study of Structural Motifs That Impact Physicochemical Properties of Relevance to Drug Discovery. *J. Med. Chem.* **2015**, *58* (22), 9041–9060.

(31) Ritchie, T. J.; Macdonald, S. J. F.; Pickett, S. D. Insights into the Impact of N- and O-Methylation on Aqueous Solubility and Lipophilicity Using Matched Molecular Pair Analysis. *Med. Chem. Commun.* **2015**, *6* (10), 1787–1797.

(32) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 354–357.

(33) Clay, K. L.; Watkins, W. D.; Murphy, R. C. Metabolism of Pyrazole. Structure Elucidation of Urinary Metabolites. *Drug Metab. Dispos. Biol. Fate Chem.* **1977**, *5* (2), 149–156.

(34) Liu, X.; Wright, M.; Hop, C. E. C. A. Rational Use of Plasma Protein and Tissue Binding Data in Drug Design: Miniperspective. *J. Med. Chem.* **2014**, *57* (20), 8238–8248.

(35) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties As a Means of Improving Compound Disposition and Safety. *Chem. Res. Toxicol.* **2011**, *24* (9), 1420–1456.

(36) Charifson, P. S.; Walters, W. P. Acidic and Basic Drugs in Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2014**, *57* (23), 9701–9717.

(37) Meanwell, N. A. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. *J. Med. Chem.* **2011**, *54* (8), 2529–2591.

(38) O'Boyle, N. M.; Boström, J.; Sayle, R. A.; Gill, A. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *J. Med. Chem.* **2014**, *57* (6), 2704–2713.

(39) Rafferty, M. F. No Denying It: Medicinal Chemistry Training Is in Big Trouble: Miniperspective. *J. Med. Chem.* **2016** *59* (24), 10859–10864.

(40) Jordan, A. M.; Begum, H.; Fairweather, E.; Fritzl, S.; Goldberg, K.; Hopkins, G. V.; Hamilton, N. M.; Lyons, A. J.; March, H. N.; Newton, R.; Small, H. F.; Vishwanath, S.; Waddell, I. D.; Waszkowycz, B.; Watson, A. J.; Ogilvie, D. J. Anilinoquinazoline Inhibitors of the RET Kinase domain—Elaboration of the 7-Position. *Bioorg. Med. Chem. Lett.* **2016**, *26* (11), 2724–2729.

(41) Newton, R.; Bowler, K. A.; Burns, E. M.; Chapman, P. J.; Fairweather, E. E.; Fritzl, S. J. R.; Goldberg, K. M.; Hamilton, N. M.; Holt, S. V.; Hopkins, G. V.; Jones, S. D.; Jordan, A. M.; Lyons, A. J.; Nikki March, H.; McDonald, N. Q.; Maguire, L. A.; Mould, D. P.; Purkiss, A. G.; Small, H. F.; Stowell, A. I. J.; Thomson, G. J.; Waddell, I. D.; Waszkowycz, B.; Watson, A. J.; Ogilvie, D. J. The Discovery of 2-Substituted Phenol Quinazolines as Potent RET Kinase Inhibitors with Improved KDR Selectivity. *Eur. J. Med. Chem.* **2016**, *112*, 20–32.

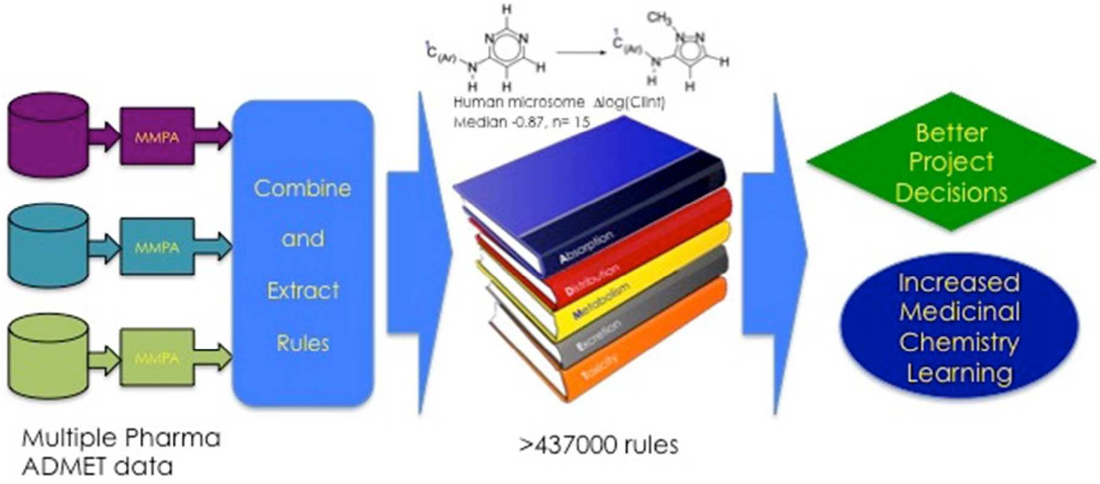
(42) Colley, H. E.; Muthana, M.; Danson, S. J.; Jackson, L. V.; Brett, M. L.; Harrison, J.; Coole, S. F.; Mason, D. P.; Jennings, L. R.; Wong, M.; Tulasi, V.; Norman, D.; Lockey, P. M.; Williams, L.; Dossetter, A. G.; Griffen, E. J.; Thompson, M. J. An Orally Bioavailable, Indole-3-Glyoxylamide Based Series of Tubulin Polymerization Inhibitors Showing Tumor Growth Inhibition in a Mouse Xenograft Model of Head and Neck Cancer. *J. Med. Chem.* **2015**, *58* (23), 9309–9333.

(43) Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* **2009**, *24* (2), 8–12.

(44) Keefer, C. E.; Chang, G.; Kauffman, G. W. Extraction of Tacit Knowledge from Large ADME Data Sets via Pairwise Analysis. *Bioorg. Med. Chem.* **2011**, *19* (12), 3739–3749.

(45) Alex, A. A.; Harris, C. J.; Smith, D. A. *Attrition in the Pharmaceutical Industry: Reasons, Implications, and Pathways Forward*; Wiley: Hoboken, New Jersey, 2016.

Table of Contents Graphic


MMPA
MMPA
MMPA
Combine
and
Extract
Rules
Multiple Pharma
ADMET data
Human microsome Δlog(Clint)
Median -0.87, n= 15
Absorption
Distribution
Metabolism
Excretion
Toxicity
>437000 rules
Better
Project
Decisions
Increased
Medicinal
Chemistry
Learning