



Deep understanding of structure–solubility relationship for a diverse set of organic compounds using matched molecular pairs

Liying Zhang^a, Hongyao Zhu^b, Alan Mathiowetz^b, Hua Gao^{b,c,*}

^a University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

^b Pfizer Research and Development, Groton, CT 06340, USA

^c Molecular Structure, Amgen Inc., 360 Binney Street, Cambridge, MA 02142, USA

ARTICLE INFO

Article history:

Received 1 June 2011

Revised 7 August 2011

Accepted 15 August 2011

Available online 22 August 2011

Keywords:

Aqueous solubility

Structural–activity relationship

Matched molecular pairs

Chemical transformation

Molecular fragmentation

Pairwise analysis

ABSTRACT

Aqueous solubility is an important biopharmaceutical property in drug discovery and development. Although it has been studied for decades, the impact on solubility by the substructures (or fragments) of compounds are still not fully understood and characterized. This study aims to obtain fragment–solubility relationships using matched molecular pairs, and to provide further insight and suggestions for chemists on structural modifications to improve solubility profiles of drug-like molecules. A set of 2794 compounds with measured intrinsic aqueous solubility ($\log S$) was fragmented into rings, linkers, and R groups using a controlled hierarchical fragmentation method. Then matched molecular pairs that differ by only one chemical transformation (i.e., addition or substitution of fragments) were identified and analyzed. The difference in solubility for each matched molecular pair was calculated, and the impact of the corresponding chemical transformation on solubility was investigated. The final product of this study was a fragment–solubility knowledgebase containing relative contributions to solubility of various medicinal chemistry design elements (R-groups, linkers, and rings). Structural modifications that might improve solubility profiles, that is, addition/deletion/substitution of fragments, could be derived from this knowledgebase. This knowledgebase could be used as an expert tool in lead optimization to improve solubility profiles of compounds, and the analysis method could be applied to study other biological and ADMET properties of organic compounds.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Biopharmaceutical properties of molecules, such as aqueous solubility, are playing a critical role in every stage of drug discovery and development. In general, poorly soluble compounds might have erratic assay results, poor bioavailability, and development challenges leading to higher probability of attrition. Meanwhile, incorrect estimation of solubility can lead to erroneous interpretation of in vitro assay results and weaken structure–activity relationship (SAR) analysis.¹

Aqueous solubility (S , mol/L) is defined as the maximum amount of solute in moles that dissolves in 1 L of water.² The logarithm of solubility, $\log S$, is widely used in solubility studies to indicate how soluble compounds are. There are two methods to determine solubility: (1) traditional thermodynamic solubility measurement, which allows a solid to equilibrate with a liquid medium, followed by sample quantification; (2) kinetic solubility measurement, in which compounds are pre-dissolved in dimethyl sulfoxide (DMSO) and the solubility is measured as the concentration at which the sample precipitates from aqueous medium. The

difference between these two solubility measurements is that kinetic solubility is not influenced by crystal lattice effects since compounds are pre-dissolved in DMSO.³

Diverse SAR analysis approaches have been used to study compounds' solubility. For example, large numbers of quantitative structure–activity relationship (QSAR) studies on solubility of organic compounds have been reported.^{4–10} Different descriptors for organic compounds have been calculated, such as physicochemical properties (e.g., partition coefficient and melting point), molecular features (e.g., molecular surface area and molecular volume), and fragments. Various machine learning methods, such as neural network, partial least square etc, have been used to build quantitative relationships between solubility and molecular descriptors. Although these QSAR models were reported to have acceptable predictive accuracies on solubility, they were highly dependent on the accuracy of descriptor calculation and usually involved model extrapolation.¹¹ Moreover, most of them could hardly provide useful structural modification suggestions for chemists on how to improve solubility because the descriptors were difficult to interpret. Meanwhile, although QSAR models based on fragment descriptors are easier to interpret, it is common for different models to have inconsistent opinions on the influence of a particular fragment on solubility, due to differences in datasets and descriptors used.

* Corresponding author. Tel.: +1(617)444 5034.

E-mail address: hgao@amgen.com (H. Gao).

Besides QSAR analysis, there is another SAR strategy that has been applied to study solubility in recent years. This approach identifies matched molecular pairs (MMPs) that differ by only a small structural mutation, i.e., chemical transformation, and calculates the corresponding changes in solubility. This approach has already been used in several publications on absorption, distribution, metabolism, excretion and toxicity (ADMET) properties.¹¹ However, most of the publications focused on one or several series of congeneric compounds. The effects on solubility due to the structural changes might not be transferable to other compound series because of limited chemical diversity. Moreover, the chemical transformations were found manually, which is an overwhelming and error prone process. Since 2006, systematically designed analysis approaches have been reported in solubility and other biological activity studies.^{11–17} Robert Sheridan et al.¹⁶ developed two SAR analysis tools, transformations analyze (T-ANALYZE), and transformation morphing (T-MORPH). By using these tools, they could automatically organize and display sets of closely related compounds for three datasets, Dopamine agonists, DHFR inhibitors for rat liver enzyme, and ACE inhibitors. However, most of the small transformation examples were only limited to several atoms, rather than functional groups. Leach et al.¹¹ identified matched molecular pairs by two sets of structural changes: the addition of substituents to an aromatic ring and the methylation of heteroatoms, as well as their effects on aqueous solubility, rat plasma protein binding, and oral exposure in rats. However, due to the methodology used, the structural changes they could analyze were limited to halogens, methyl and other small terminal substituents of aromatic compounds. Lewis et al.¹⁴ used Pfizer in-house data mining tool (known as 'Buy me Grease') to evaluate how specific alterations on chemical structure can affect in vitro ADME properties over 150,000 human liver microsome (HLM) clearance data points. Again, the structure changes were only small substituents on phenyl rings. All of these reports provided examples with small structural changes, which technically could be found manually. The reports lacked studies on relatively large and diverse structural changes, such as a change of rings or linkers, which could have profound effects on solubility as well.

In this work, we studied the effects of functional groups/fragments, such as rings, linkers, and R-groups, on solubility. After fragmenting compounds using a controlled hierarchical fragmentation algorithm, we developed an automatic way to find MMPs that differ by only a particular chemical transformation (i.e., addition or substitution of fragments). The differences in solubility for all MMPs were calculated. This results in a set of structural rules that describe the relative solubility contributions of different medicinal chemistry design elements (R-groups, linkers, and rings). This is the first study, to our knowledge, of chemical transformation in terms of functional groups, which might be easier for chemists to understand and apply in lead optimization. The resulting knowledgebase is being used as an expert in silico solubility tool at Pfizer during lead optimization.

2. Methods

2.1. Solubility dataset

A set of 2794 compounds with measured intrinsic solubility data was used in this study. Most of the solubility data were measured at Pfizer using equilibration time of 24 h or longer. 720 solubility data points were obtained from the literature.⁷ Figure 1 shows the distribution of logS in this dataset. The maximum and minimum solubility are 1.60 and –7.98 log units. Usually, experimental solubility data provided an average standard deviation of 0.6 units.¹¹ The data is skewed since there is an upper limit to the solubility.

2.2. Hierarchical molecular fragmentation approach

The hierarchical molecular fragmentation algorithm was developed based on the fragmentation module imbedded in Molecular Operating Environment (MOE, 2009.03)¹⁸ using the Scientific Vector Language (SVL).¹⁹ The fragmentation of compounds with any rings (a.k.a. cyclic compounds) proceeds as depicted in Figure 2: (1) Level 1: removing terminal side chains, or R-groups to obtain the molecular framework. Exocyclic double bonds and double bonds directly attached to the linkers are kept²⁰; (2) Level 2: removing linkers to obtain the scaffolds of molecules. Rings that are connected through a single bond are kept. Atoms that are directly connected to rings through double bond are kept; (3) Level 3: separating those rings that are directly connected to obtain the ring fragments of molecules. Fused rings are kept. After these three steps, compounds were fragmented into rings, linkers and R-groups, which were represented and stored by SMARTS. Meanwhile, the connection points of each fragment to others were well labeled. The information of whether a fragment was connected to rings or not was also stored in the SMARTS strings.

2.3. Matched molecular pairs (MMPs)

In this study, MMP is defined as a pair of molecules that differ only by a single chemical transformation (among rings, linkers, and R-groups). There are two types of chemical transformation studies in this work: (1) type I chemical transformation: addition/deletion of a fragment (R-group). If after replacing a fragment in a molecule by hydrogen, the molecule is identical with another molecule, these two molecules are considered a type I MMP. This addition/deletion of fragment is considered as a chemical transformation as well since it is the substitution between hydrogen and a fragment; (2) type II chemical transformation: mutation of a fragment to another fragment. In this study, the fragments were confined in the substructures generated by hierarchical fragmentation approach. In other words, the fragments in this study are defined as rings, linkers or R-groups derived from the solubility dataset. In order to collect as many/diverse transformations as possible, there is no structural limitation for the mutation. In other words, as long as these two fragments are not structurally identical, we consider them as a type II chemical transformation. For example, if after one fragment α in molecule A is changed to another fragment β , A is identical with another molecule B, then A and B are considered as a type II MMP.

MMPs were searched through the entire dataset automatically using in-house scripts encoded using SVL language in MOE. Since the fragments used for searching were defined by SMARTS, the structures were tightly restricted. There are two schemes used in identifying MMPs and the results from these two schemes complement to each other: (1) if a fragment (R-group) is given, we could identify type I MMPs using in-house script, `one_group_less`. This script first deletes this fragment from compound B, to form compound C, and then searches the entire dataset to find compounds identical with C. If compound A is found, then B and A are defined as one pair. The change of corresponding solubility from A to B due to the addition of this R-group will be calculated as the following:

$$\Delta \log S = \log S_B - \log S_A \quad (1)$$

A positive $\Delta \log S$ indicates that the addition of this fragment increases solubility, while a negative value means the deletion of this fragment could result in more soluble compound.

It is expected that for each fragment, there will be more than one MMPs found in the dataset. Therefore after all the pairs for this fragment were found, the median $\Delta \log S$ were calculated, along with corresponding standard deviation and standard error of the mean.

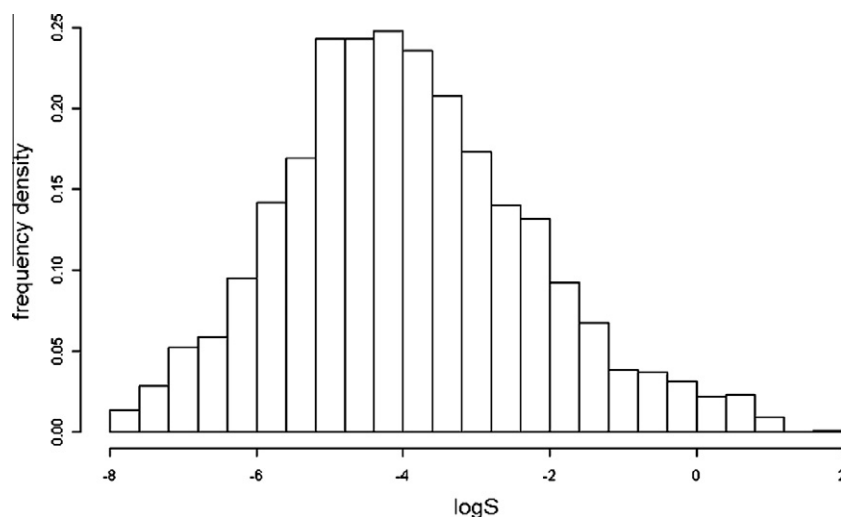


Figure 1. The distribution of $\log S$ of the solubility dataset for this study.

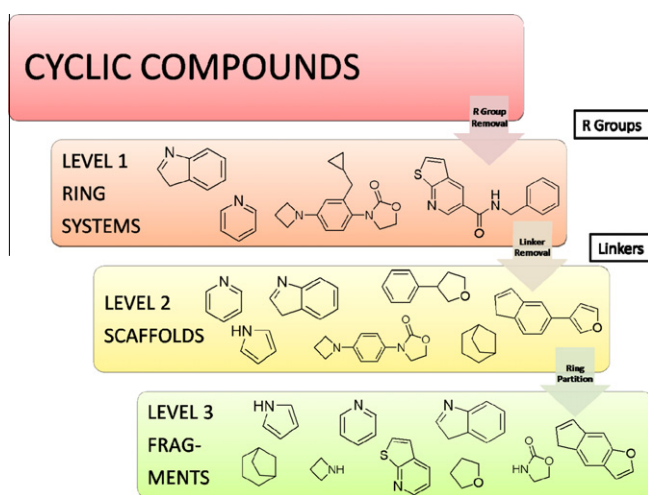


Figure 2. Hierarchical molecular fragmentation algorithm.

If no fragment is specified, we will identify type I and II MMPs by in-house script `one_group_diff`. For any two compounds A and B, this script first fragments the compounds into R-groups, rings and linkers. Then it compares all the fragments of A and B, and considers them as a MMP if there is only one fragment difference between them. This chemical transformation could be either addition/deletion or substitution of a fragment. During the comparisons, the connection points and connection order of fragments were taken into consideration, in order to filter out ‘false positive’ pairs.

The change of corresponding solubility due to the chemical transformation was calculated using Eq. 1. Again a positive $\Delta \log S$ indicates that the transformation (from A to B) could increase solubility. After all MMPs were found, we collected the pairs with same chemical transformation and calculated median $\Delta \log S$, as well as corresponding standard deviation and standard error of the mean.

2.4. Crystal-packing motifs

Crystal-packing motifs are the functional groups in compounds that can form intermolecular hydrogen bonds, which could increase the melting point of compounds and decrease equilibrium

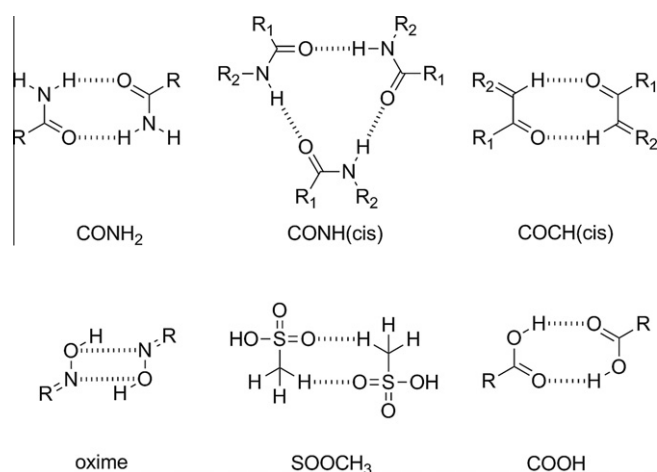


Figure 3. Examples of crystal-packing motifs. Dashed lines are hydrogen bonds.

solubility. Those motif interactions, that is, hydrogen bonds, can be formed homomerically or heteromerically. Some common crystal-packing motif examples are shown in Figure 3.²³ The hydrogen bonds are shown in dashed lines. In this study, we searched MMPs whose chemical transformations can form crystal-packing motifs, such as amide and carboxyl groups. The results were analyzed along with other fragments.

3. Results and discussion

We collected all matched pairs found in the solubility dataset of 2794 drug-like molecules using in-house scripts. Two molecules in a matched pair differ by only a particular chemical transformation, that is, addition/deletion or substitution of fragments. The fragments could be rings, linkers connecting two rings, or R-groups connected to rings. These fragments were generated by hierarchical fragmentation approach. Then the corresponding change in solubility ($\Delta \log S$) was calculated for each MMP. For each chemical transformation, it was expected that more than one MMP could be found in the dataset. Therefore, after all MMPs were found, we collected the pairs with same chemical transformation and calculated median $\Delta \log S$ and other statistical parameters.

MMPs found in this study were separated into two types according to the differences of the corresponding chemical transformations

Table 1
Examples of type I and II matched molecular pairs

	Structure A		Structure B	$\Delta \log S$
Type I				
	 logS: -0.14		 logS: -1.82	-1.68
Type II	 logS: -2.84		 logS: -1.95	0.89
	 logS: -5.69		 logS: -3.52	2.17
	 logS: -2.14		 logS: -3.05	-0.91

(Table 1): (1) the chemical transformation of type I MMPs is the addition/deletion of a fragment. Most of the fragments are R-groups connected to rings; (2) the chemical transformation of type II MMPs is the substitution/mutation of a fragment. The fragments could be rings, linkers or R-groups. In this work we have identified 1920 type I MMPs and 4196 type II MMPs. Their impacts on aqueous solubility will be discussed later in this section.






















3.1. Type I matched molecular pairs

If after replacing a fragment in a molecule by hydrogen, the molecule is identical with another molecule, then we consider these two molecules as a type I MMP. We fragmented the chemical structures in the solubility dataset and identified 42 R-groups. Then by searching through the dataset using both schemes mentioned above, 1920 type I MMPs were found for these R-groups. The solubility of both compounds in each pair was extracted from database, and the difference of solubility ($\Delta \log S$) was calculated using Eq. 1. For each chemical transformation, all matched pairs were then collected and the median $\Delta \log S$ was computed. Positive $\Delta \log S$ indicates that the presence of this fragment could increase solubility. The percentage of MMPs in which the presence of this fragment increased solubility was calculated as well. Previous work suggested that at least 20 pairs for one chemical transformation is needed to provide statistical significant conclusions in SAR analysis.²² In total 12 R-groups were derived with more than 20 MMPs found in the dataset. Due to the diversity of the dataset, many chemical transformations were found with fewer than 20 MMPs pairs. However, the effects of these chemical transformations on solubility might be of great interest as well. Therefore, a total of 26 R-groups were collected and listed in Table 2. For many of these R-groups, the $\Delta \log S$ of MMPs were in a wide range. Therefore median $\Delta \log S$ was calculated for every R-group to provide a general impression of its impact on solubility. Other statistical parameters, such as the standard deviation and standard error of the mean, were also calculated and listed.

It needs to be pointed out that, ideally, differences in base structures need to be considered when calculate the general effect on solubility (median $\Delta \log S$) for a particular chemical transformation. Especially for small R-groups, for example, halogens, the effect of base structures might be the major reason for the broad range of $\Delta \log S$. However, if base structures are considered, the number of MMPs per base structure per fragment would dramatically decrease, and thus makes it even harder to generate statistical significant conclusions. Therefore in this study, the differences and impacts of base structures are not included in the discussion, while their importance is fully understood. Actually, this is a common solution when studying chemical transformations in diverse datasets.¹²

Some expected effects on solubility of these chemical transformations were observed in Table 2, such as halogens. Our study showed that the addition of fluorine, chlorine, and bromide could decrease solubility (Table 2, Lines 1–4). This observation has been reported in many SAR analyses on solubility^{11,17} and widely applied by chemists. Additionally in our study, the median $\Delta \log S$ for halogen substitution could reveal how large their impacts on solubility are. For example, compounds with chlorine, on average, could decrease solubility by 1.45 log units, when compared with compounds without chlorine at the same position of the structures, and 96% of 116 pairs have this trend. Meanwhile, the MMPs for fluorine have a higher median $\Delta \log S$, and the matched pairs for iodine have a lower $\Delta \log S$. In other words, the heavier halogen has lower $\Delta \log S$. The decreasing trend in $\Delta \log S$ for halogens correlates well with their electronegativity and hydrophobicity. This close correlation between $\Delta \log S$ and electronegativity was also seen in halogen-containing fragments (Table 2, Line 5): Trifluoromethyl is often described to have intermediate electronegativity between fluorine and chlorine,²⁴ the same trend was observed when comparing its $\Delta \log S$ with fluorine and chlorine. Although there are only 12 pairs found for trifluoromethyl, its effect on solubility in our study is consistent with other SAR reports.¹⁶ According to these findings, adding halogens most likely could decrease aqueous solubility.

Table 2Effects of 26 common R-groups on solubility in type I matched molecular pairs^a.

	Fragment	Structure	#Matched pairs	Median $\Delta\log S$	Max $\Delta\log S$	Min $\Delta\log S$	SD	SEM	%Pairs increase solubility
1	F[R]	R-F	50	-0.45	0.77	-2.16	0.68	0.10	22%
2	Cl[R]	R-Cl	116	-1.45	0.40	-4.06	1.03	0.10	4%
3	Br[R]	R-Br	13	-0.72	0.15	-2.13	0.64	0.18	15%
4	I[R]	R-I	4	-2.07	-1.00	-3.32	0.97	0.48	0%
5	FC(F)([R])F		12	-0.77	0.35	-1.60	0.69	0.20	25%
6	C[R]	R-	193	-0.50	2.30	-2.74	0.81	0.06	26%
7	CC[R]		65	-1.12	0.82	-3.03	0.72	0.09	9%
8	CC([R])C		9	-1.40	0.14	-2.99	0.85	0.28	11%
9	CCC[R]		39	-1.72	0.76	-3.38	0.77	0.12	5%
10	CC(C)([R])C		3	-1.84	-1.83	-2.26	0.25	0.14	0%
11	CCCC[R]		30	-2.34	-0.35	-3.08	0.72	0.13	0%
12	[H]OC[R]		49	0.61	2.12	-0.58	0.58	0.08	78%
13	[H]OCC[R]		22	0.66	1.98	-0.87	0.81	0.17	64%
14	[H]OC(C)([R])C		4	0.92	1.09	0.62	0.21	0.10	100%
15	CO[R]		42	-0.24	0.73	-1.59	0.62	0.10	43%
16	CS[R]		7	-0.81	-0.13	-1.28	0.42	0.16	0%
17	O=C([R])N([H])[H]		9	-0.86	1.02	-2.17	1.04	0.35	33%
18	O=C([R])N([H])C		5	-0.84	0.52	-0.97	0.74	0.33	40%
19	O=C(C)N([H])[R]		2	0.22	0.62	-0.18	0.57	0.40	50%
20	[H]OC([R])=O		20	-0.05	1.90	-1.64	1.02	0.23	45%
21	[O-][S+2]([O-])([R])C		8	-0.38	0.00	-2.10	0.68	0.24	0%
22	[O-][S+2]([O-])([R])O[H]		3	-0.94	-0.42	-1.12	0.36	0.21	0%
23	[O-][S+2]([O-])([R])N([H])[H]		3	-0.40	-0.32	-1.13	0.45	0.26	0%
24	CO[P+](OC)(O[R])[S-]		3	-1.82	-1.76	-2.21	0.25	0.14	0%
25	[H]N([R])[H]		23	0.76	2.67	-0.9	0.97	0.17	61%
26	[H]O[R]		71	0.97	3.32	-0.88	1.07	0.13	85%

^a SD, Standard Deviation; SEM, Standard Error of Mean; R indicates connection point.

It is well known that in general alkyl groups could decrease solubility due to their lipophilic nature. We then studied the relationships between different alkyl groups and their effects on aqueous solubility (Table 2, Lines 6–11). Based on the median $\Delta\log S$ values

for several alkyl groups (from methyl group to butyl group), we found that the longer alkyl groups, the lower the solubility of the compounds. This observation is consistent with previous reports.²¹ Therefore we could improve molecule solubility by decreasing the

length/size of alkyl groups. Addition of hydroxyl group could compensate the decreasing effects on solubility by alkyl groups (Table 2, Lines 12–14).

Another interesting finding in the results is that not all polar groups can increase solubility. A very typical example is amide (Table 2, Lines 17–19). Primary amide and secondary amide, as well as reverse amide all decreased solubility in 63% of the MMPs. Another example is the carboxyl group, which decreased solubility in 55% of the cases. Furthermore, the presence of sulphonamide and sulfonyl groups always decreased solubility in the 17 pairs found in our studies (Table 2, Lines 21–24). There are many possible reasons to explain why the incorporation of these polar groups can cause a decrease in solubility. For example, many physicochemical properties of the entire or the partial structures, for example, pK_a , lipophilicity, zwitterions, could have critical impacts on the solubility (when study the compounds with and without w/o polar groups). However, such discussion is more suitable for SAR analysis on congeneric series of compounds since it requires the examination of each structure. Due to the size and diversity of the dataset we studied, we decided to focus on the possible reasons that are usually neglected and might be more appealing to chemists. A plausible explanation is that all of these groups are crystal-packing motifs. Crystal-packing motifs could form intermolecular hydrogen bonds and other interactions and result in thermodynamically stable crystalline. This leads to an increased melting point and decreased solubility. Amides, carboxyl and sulfonyl groups have been considered as potent crystal-packing motifs.²³ This might explain why more than half of the molecules containing these polar fragments were found to have lower solubility than their matched molecules.

The solubility dataset in this study is measured equilibrium solubility, thus the solubility of compounds could be highly influenced by crystal-packing effect. In order to prove that crystal-packing motifs indeed play an important role in decreasing intrinsic solubility, we analyzed another dataset of more than 20,000 compounds, of which the solubility was measured kinetic solubility. Kinetic solubility is not influenced by crystal-packing effects due to its measurement (details in Section 2). Therefore we examined and compared the MMPs for these crystal-packing motifs in both datasets. Take the carboxyl group for example, half of the MMPs had decreased solubility in equilibrium dataset, while in kinetic dataset almost all MMPs increased or kept the solubility (Fig. 4).

For these 42 R-groups collected from the solubility dataset, we compared its Hansch π constant and median $\Delta\log S$ (Fig. 5). Hansch π constant measures the hydrophobicity of fragments. The higher Hansch constant is, the more hydrophobic the fragment is. Ideally Hansch π constant and $\Delta\log S$ should be negatively correlated. Most R-groups we studied showed this trend in Figure 5. There are 5 outlier fragments, all of which are amide groups and sulfonyl groups ([OH][S+2]([O-])([O-]), [CH₃][S+2]([O-])([O-]), [NH₂][S+2]([O-])([O-]), [NH₂]C(=O), [NH](C(=O))[CH₃]). They are potential crystal-packing motifs and might help the compounds to form stable crystalline, which could decrease their solubility. This might be the reason why their median $\Delta\log S$ are lower than expected.

3.2. Type II matched molecular pairs

If after replacing one group α in molecule A with another group β , A is identical with another molecule B, A and B are considered as a type II MMP. The solubility of both compounds was extracted from database, and the difference of solubility ($\Delta\log S$) was calculated. A positive $\Delta\log S$ indicates that replacing group β by group α could increase solubility. For each chemical transformation, all matched pairs were then collected and the median $\Delta\log S$ was computed and analyzed.

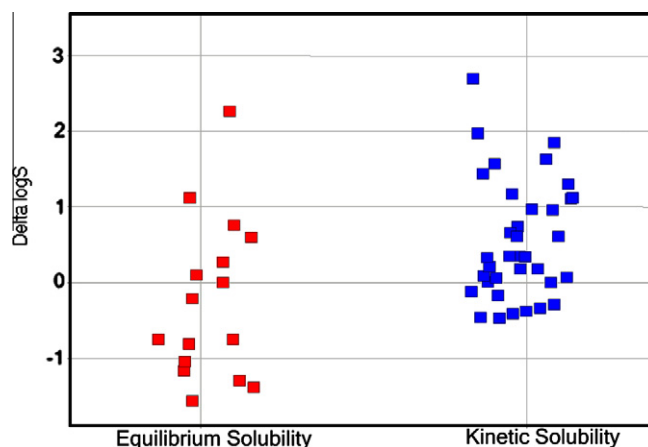


Figure 4. Comparison of $\Delta\log S$ of MMPs for carboxyl group in different solubility datasets. Red cubic: equilibrium solubility dataset, which is influenced by crystal-packing effects; blue cubic: kinetic solubility dataset, which is not influenced by crystal-packing effects.

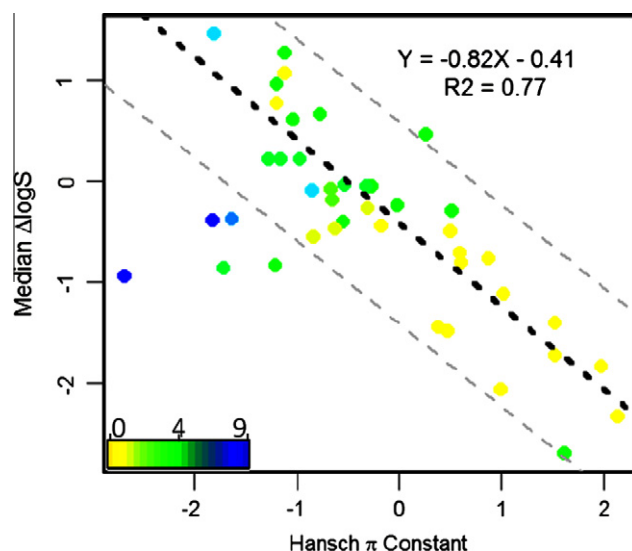


Figure 5. Hansch π constant vs. median $\Delta\log S$ for 42 R-groups. Color of the point represents the count of H-bond donor/acceptors of the R-group (minimum 0 and maximum 9). The linear regression (black dash line) equation for the 37 points (5 outliers were excluded) is on the top right corner of the plot. Grey dash lines represent the deviation of 1 unit of $\Delta\log S$.

In this solubility dataset, 4196 type II MMPs were found for 3397 chemical transformations. Technically the chemical transformation could happen between any types of fragments, such as ring-to-linker, linker-to-R-group, etc. In this study, we focused on ring-to-ring, linker-to-linker, and R-to-R transformations, since such transformations are of most interests for chemists during lead optimization. It is obvious that in the real practice of structure modification, chemists are prone to modify R-groups. This is why more than half of the transformations we found are for R-groups. In total we have found 112 ring-to-ring, 22 linker-to-linker, and 141 R-to-R transformations. It is an interesting but neglected question of how ring or linker modifications could influence compounds' physicochemical and biological properties, and our analysis strategy is a useful tool for such studies.

For each transformation, we collected all MMPs and calculated the median $\Delta\log S$. Heat maps were then generated indicating the effects that each transformation could have on solubility (Fig. 6).

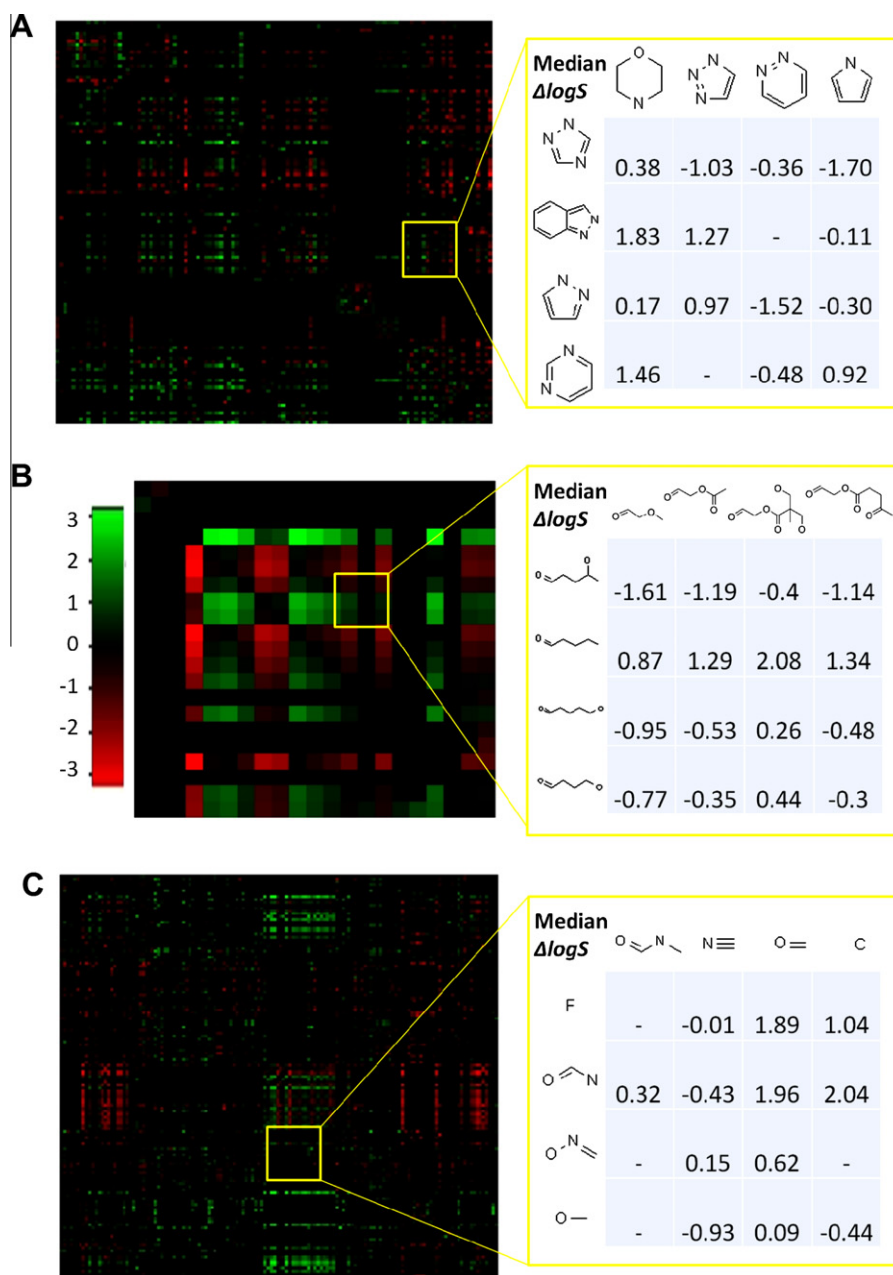


Figure 6. Heat maps for the contributions of type II chemical transformations on solubility. (A): ring-to-ring transformation; (B): linker-to-linker transformation; (C): R-to-R transformation. Green indicates positive $\Delta\log S$ and red indicate negative $\Delta\log S$. Tables on the right are examples from the heat maps. Transformations start from the fragments in the first column, to the fragments in the first row. The yellow highlighted squares on heat maps do not necessarily represent the area from which the examples were extracted.

The heat maps contain three parts: (1) Scores for ring transformations, indicating the average change in solubility where a given ring structure is replaced with another ring structure; (2) Scores for linker transformations, indicating the average change in solubility for compounds where a given linker is replaced with another; (3) Scores for R-group transformations, indicating the average solubility change when changing one R-group to another in the context of an aliphatic or aromatic ring system. The structural information for each fragment and the statistics of $\Delta\log S$ for each chemical transformation could be found in [Supplementary data](#).

Figure 6 also shows some examples of the usage of heat maps. Assuming chemists would like to improve the solubility of a lead compound A, they decide to modify the pyrazole group within A,

since they know this fragment does not have a major impact on its biological activity. Therefore by looking up the heat maps (heat map A since this is a ring group, or the excel file in the [Supplementary data](#)), a set of chemical transformations that start with pyrazole group could be identified (right top table in Fig. 6, third row). After comparing the possible impacts on solubility ($\Delta\log S$) of these chemical transformations, they could decide to substitute the pyrazole with a triazole group.

The heat maps could be considered as a knowledgebase describing the relative solubility contributions of different medicinal chemistry design elements (R-groups, linkers, and rings) since they record the effects of chemical transformations on aqueous solubility. It would be of great values for chemists to get suggestions on structural modification in order to enhance solubility.

4. Conclusions

Many big pharmaceutical companies have built up rich SAR chemical libraries, which contain an abundance of information on how chemical transformations could influence biopharmaceutical, ADMET, as well as biological properties of drug-like molecules. In this study, we described a systematic approach to identify MMPs in a diverse solubility dataset, studied fragment–solubility relationships, and demonstrated how this can be used to improve the solubility of compounds. It needs to be mentioned that the intention of this investigation was to build a knowledgebase for fragment–solubility relationship. Due to the diversity, the size of the dataset is still relatively small, and we thus did not separate the dataset into training and test sets as QSAR analysis commonly does. More quantitative fragment–solubility relationship analysis is beyond the scope of this study.

A set of 2794 compounds with intrinsic aqueous solubility data was fragmented into linkers, rings and R-groups, by a hierarchical fragmentation method. MMPs that differ only by a particular chemical transformation (i.e., addition or substitution of rings, linkers or R-groups) were compiled. The difference in solubility ($\Delta\log S$) for all MMPs with same chemical transformation was computed, and the effect on solubility of each chemical transformation was then studied based on the median $\Delta\log S$. Suggestions of structural modification, that is, addition/deletion/substitution of fragments, are then provided in order to enhance the solubility in chemical design. In this study, we have analyzed the effects upon solubility for many chemical transformations, and built a knowledgebase for aqueous solubility. This approach could be applied to study and improve other biological activities of chemicals.

It needs to be pointed out again that the chemical transformations we studied are functional groups and fragments that are commonly accepted by medical chemists as design elements during lead optimization. Compared with previous chemical transformation studies (not only for solubility),^{11–17} such chemical transformations and their impacts on physic-chemical/biological activities would be more interesting and acceptable for chemists. For example, chemists would treat the change from benzene to pyridine as a ring transformation. However, according to some previous chemical transformation studies, this could be considered as an atom transformation (from aromatic carbon to nitrogen). Another interesting and novel contribution of this study is that we listed the impacts of ring and linker fragments to solubility, while a number of previous studies focused only on the impacts of R-groups. Generally, it would be much harder for chemists to identify the impacts of ring and linker than R group transformations.

Therefore we believe that our analysis would be an interesting and valuable tool for medicinal chemists.

An 'Expert Solubility Systems' has been developed to incorporate the findings of this study. By using this tool, a given compound with poor solubility will be fragmented into R-groups, linkers, and ring systems first. Then by looking up the knowledgebase, possible structural motifs (rings, linkers, or R-groups) will be suggested to modify the original structures to improve solubility. Naturally, these suggested modifications need to be weighed with other considerations, such as certain structural requirements to maintain biological activity or favorable ADMET properties.

Acknowledgment

This study was supported by Pfizer summer internship research program.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bmc.2011.08.036](https://doi.org/10.1016/j.bmc.2011.08.036).

References and notes

1. Faller, B.; Ertl, P. *Adv. Drug Deliv. Rev.* **2007**, *59*, 533.
2. Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. *Curr. Med. Chem.* **2006**, *13*, 223.
3. Dehring, K. A.; Workman, H. L.; Miller, K. D.; Mandagere, A.; Poole, S. K. *J. Pharm. Biomed. Anal.* **2004**, *36*, 447456.
4. Chen, X. Q.; Cho, S. J.; Li, Y.; Venkatesh, S. *J. Pharm. Sci.* **2002**, *91*, 1838.
5. Delaney, J. S. *Drug Discovery Today* **2005**, *10*, 289.
6. Eros, D.; Keri, G.; Kovesdi, I.; Szantai-Kis, C.; Meszaros, G.; Orfi, L. *Mini. Rev. Med. Chem.* **2004**, *4*, 167.
7. Gao, H.; Shanmugasundaram, V.; Lee, P. *Pharm. Res.* **2002**, *19*, 497.
8. Huuskonen, J. *Environ. Toxicol. Chem.* **2001**, *20*, 491.
9. Wanchana, S.; Yamashita, F.; Hashida, M. *Pharmazie* **2002**, *57*, 127.
10. Xia, X.; Maliski, E.; Cheetham, J.; Poppe, L. *Pharm. Res.* **2003**, *20*, 1634.
11. 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.
12. Hajduk, P. J.; Sauer, D. R. *J. Med. Chem.* **2008**, *51*, 553.
13. Hussain, J.; Rea, C. *J. Chem. Inf. Model.* **2010**, *50*, 339.
14. Lewis, M. L.; Cucurull-Sanchez, L. *J. Comput. Aided Mol. Des.* **2009**, *23*, 97.
15. Papadatos, G.; Alkarouri, M.; Gillet, V. J.; Willett, P.; Kadirkamanathan, V.; Luscombe, C. N.; Bravi, G.; Richmond, N. J.; Pickett, S. D.; Hussain, J.; Pritchard, J. M.; Cooper, A. W. J.; Macdonald, S. J. *J. Chem. Inf. Model.* **2010**, *50*, 1872.
16. Sheridan, R. P.; Hunt, P.; Culbertson, J. C. *J. Chem. Inf. Model.* **2006**, *46*, 180.
17. Warner, D. J.; Griffen, E. J.; St-Gallay, S. A. *J. Chem. Inf. Model.* **2010**, *50*, 1350.
18. Molecular Operating Environment. 2009. Chemical Computing Group.
19. Scientific Vector Language. 2009. Chemical Computing Group.
20. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *J. Chem. Inf. Model.* **2007**, *47*, 47.
21. Forster, S.; Buckton, G.; Beezer, A. E. *Int. J. Pharm.* **1991**, *72*, 29.
22. Clark, M. *J. Chem. Inf. Model.* **2005**, *45*, 30.
23. Mercury 2.3. 2009. Cambridge Structural Database System.
24. True, J. E.; Thomas, T. D.; Winter, R. W.; Gard, G. L. *Inorg. Chem.* **2003**, *42*, 4437.