## ARTICLES

# A Database of Historically-Observed Chemical Replacements

David Y. Haubertin and Pierre Bruneau*

AstraZeneca, Centre de Recherches, Z.I. La Pompelle, BP 1050, Chemin de Vrilly, 51689 Reims, Cedex 2, France

Received September 12, 2006

A systematic analysis of one-to-one chemical replacements occurring in a set of 50 000 druglike molecules was performed. The frequency of occurrence, as well as the average change in measured and calculated properties, was computed for each observed substitution. The experimental properties considered were solubility, protein binding, and logD. The calculated properties were logP, molecular weight, number of hydrogen bond donors and acceptors, and polar surface area. During this analysis, in which 9000 different functional groups were considered, 0.7 million substitutions were identified and stored in a database. As an application, we present a web interface from which users can identify historically observed replacements of any functional group on their query molecule. The server returns a list of side-chains, as well as the historically observed shift in experimental properties.

## INTRODUCTION

At every stage in the drug discovery process, medicinal chemists have to choose which compounds to synthesize. Virtual chemistry space is large and might contain $10^{100}$ molecules,[1] a number far too large to be explored systematically. Even though technologies such as combinatorial chemistry and high-throughput screening (HTS) offer ways to explore chemistry space faster, only a fraction of that space can be explored.

Depending on the stage in the drug discovery process, an exhaustive exploration might not necessarily be the goal of the chemist. To generate a lead, *hit identification* and *lead generation* phases require screening of numerous regions of chemistry space. The *lead optimization* phase only requires an exploration of a particular region of the chemistry space, the center of which would be the previous lead.

Modifications are brought to that lead to improve certain physicochemical or biological properties and move the initial compound to the status of *drug candidate*, therefore exploring the chemistry space surrounding the initial lead. In that stage, modifications mostly include side-chain replacements with a change of scaffolds being rare. Models can be made to predict any given property, but it is much more difficult to predict what modifications should be made to the compound to increase or decrease that property. Obvious substitutions are those governed by Lipinski's rule-of-five[2] to improve absorption and permeation. To increase permeability, a chemist might for example decide to substitute a chlorine atom with a methyl group to lower the molecular weight and fall below the 500 g $mol^{-1}$ barrier or render the molecule more lipophilic so that it can cross gastrointestinal membranes more easily. These substitutions are quite easy to comprehend and can be directly linked to side-chain properties.

The same cannot be said of other properties that medicinal chemists try to fine-tune in late-stage drug discovery projects. Solubility is more difficult to understand. It is the result of a complex interplay of several factors including the hydrogen-bonding nature of the compound and the energetic cost of disrupting the crystal lattice of the solid to bring it into solution.[3] In that case, it becomes more difficult for a chemist to predict accurately what replacement to make to improve solubility. Very few methodologies dealing with inverse QSAR[4,5], that is, finding out what changes are needed to render a molecule more active or soluble, have been reported in the literature. Models exists that predict solubility,[6] but they only give a global answer and do not help in the selection of substitutions that might help improve that property. The same can be said of protein-binding or pharmacokinetic properties for example.[7,8] In those cases, by experience, a chemist might suggest a particular substitution, knowing that in another project or in the literature, the same substitution was associated with an increase or decrease in any given property. While this knowledge-based approach might be more successful than in silico models in predicting the effect of replacements, it is limited to the amount of knowledge of the chemist and is highly biased toward their own experience and culture. Other substitutions might have been observed in other projects that might be just as good but that did not happen to have come to the attention of the chemist. Pharmaceutical companies all store information relating to their compounds collection in databases; it is fairly quick and easy to retrieve all the available data for a given molecule. However, information concerning the properties of not one molecule but of a pair of molecules represents a wealth of data that is only partially exploited. A systematic method to handle this information and to find substitutions

* To whom correspondence should be addressed. E-mail: pierre.bruneau@astrazeneca.com.

10.1021/ci600395u CCC: $37.00 
Published on Web 06/01/2007

and their associated changes in physicochemical properties would generalize the concept of knowledge-based modification of a molecule. This would allow medicinal chemists to explore regions of chemical space that they would not have explored solely on the basis of their own experience or knowledge, by identifying un-noticed substitutions.

There are several side-chain studies reported in the literature. Bemis and Murcko examined the distribution of side-chains in drug molecules[9] and discussed possible applications to molecular design, arguing that molecules generated using these side-chains are more likely to be druglike compounds. Lewell[10] and Ertl[11] went a bit further and characterized the drug likeliness of substituents. Side-chains with a relatively higher representation in drugs than in nondrugs were given a higher likeliness index. Ertl[11] describes the application of such a study to a web tool for the automatic identification of bioisosteric substituents. Bioisosterism[12] can be defined as the replacement of one functional group by another in a bioactive molecule that retains biological activity. Sheridan[13] extracted one-to-one replacements of chemical groups in pairs of druglike molecules with the same biological activity and counted the frequency of the replacements in the MDDR database.[14] In contrast to studies where bioisosterism was examined, our goal was to identify replacements that would increase or decrease a given property. Solubility at pH 7.4, logD at pH 7.4, serum protein binding, and calculated properties (molecular weight, ClogP, etc.) were the targeted properties. ACD/Structure Design Suite[15,16] is a tool that has been developed by Advanced Chemistry Development (ACD) to suggest analogs with appropriate modifications to improve related ADME properties, including solubility and logD. Whereas we aim to identify historically observed experimental shifts in properties, the ACD software predicts shifts based on calculated properties.

Existing methods are either static studies, focused on bioisosterism, or are based on calculated/predicted properties. They do not answer the following question: What chemical replacements have been historically associated with an increase in solubility and a decrease in protein binding? This paper describes our efforts to address this issue. The objective was to analyze and extract side-chain occurrences in molecules from our corporate databases and to document the knowledge gained in a web-searchable format. The paper is laid out as follows. After describing the principles of the system, we will describe the datasets that were used to build the database; subsequently, we will discuss the methodology used to build the database. We will then describe the database and the graphical user interface developed. Finally, possible applications of the database will be discussed.

## PRINCIPLES

The main idea behind this work is to help the chemist find novel molecules while improving certain physicochemical properties, that is, suggest replacements for side-chains that satisfy certain conditions. "Substituents" or "side-chains" are defined in the fashion outlined by Ertl[11] as any group of atoms connected by a single chemically activated (breakable) bond to the rest of the molecule. Our system basically answers the following question: given a molecule, with what should I substitute this side-chain with to increase/decrease this property? Seen as a black box, the entry is a molecule and a few filters and the output is a list of molecules with their associated predicted change in properties. The system we developed involves four components that interact with each other. The core of the system is the substitution database that holds historical information and has been set up as a MySQL database. The second part of the system is the user interface, implemented as a web server. It allows the user to query the database interactively. The third component is an algorithm that decomposes the user molecule, therefore enabling side-chain detection. The last component is a list of side-chains. A summary of the process, seen from the user side, is described in Figure 1. All compounds and fragments are internally described using the SMILES notation.[17] This allows an easy manipulation of data and a compact way of storing structures. The user first feeds in a query molecule (section 1), either by drawing it or entering its SMILES notation. The compound is then decomposed into side-chain, remainder pairs by spotting side-chains from the database that are present in the molecule (section 2). The user is then presented with these combinations, picks the ones of interest, and selects filters (increase/decrease/range, in any given property) (section 3). Once the choices have been validated, the system performs an SQL query on the substitution database (the construction of which will be detailed later) using the selected filters and side-chains (section 4). Finally, the system returns the list of side-chains and built molecules to the user (section 5). In summary, with an input compound and some limits on physicochemical properties, our system suggests new compounds, each of them corresponding to the replacement of one side-chain in the input molecule.

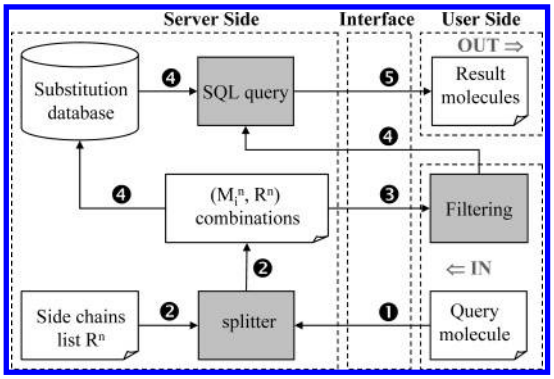



**Figure 1.** Schematic representation of the server user interface. The user feeds in a query molecule (1). The molecule is split into a list of core and side-chain combinations (2). The list is returned to the user (3), who selects the combinations of interest and desired range of increase/decrease of a property/properties. Selected side-chains and filters provided by the user are used to query the substitution database (4) and return a list of suggested molecules (5).

## DATA SET

**Compound Dataset.** We initially limited this study to the compounds available within AstraZeneca. Other data were considered from sources like the MDL Drug Data Report (MDDR) dataset,[14] but the heterogeneity of such a set would have involved complex preparation work to render the data homogeneous. The use of in-house data ensures us of (i) using druglike molecules and (ii) using experimental properties with defined protocols and identical method over the

**Table 1.** Average Physicochemical Properties of Druglike Molecules Used in the Database

| | $\mu$ | $\sigma$ | median | $N$ |
|---|---|---|---|---|
| log solubility (M) | −4.61 | 1.23 | −4.56 | 34 484 |
| log $K1_{app}$ | 4.80 | 0.88 | 4.77 | 12 651 |
| logD | 2.26 | 1.24 | 2.39 | 27 838 |
| ClogP | 3.71 | 1.62 | 3.70 | 54 427 |
| MW (g $mol^{-1}$) | 431.97 | 95.74 | 435.3 | 54 427 |
| HBD | 1.82 | 1.40 | 2 | 54 427 |
| HBA | 6.58 | 2.33 | 6 | 54 427 |
| PSA (Å$^2$) | 86.43 | 33.11 | 85.26 | 54 427 |
| rule-of-5 | 0.48 | 0.71 | 0 | 54 427 |

entire set. The methodology presented in this paper is nevertheless easily portable to other datasets and experimental properties.

Properties we considered were partly experimentally measured and partly calculated descriptors. Measured properties included solubility, logD, and rat serum protein binding. These properties were chosen because of the large pool of data that was available in-house. This would not have been the case if we had considered pharmacokinetic data for example.

The choice of these criteria also ensured examination of only druglike compounds because the characterization of solubilities, logD, and protein binding are not carried out for intermediates.

Solubility data are derived from an experiment in which solid compound is added to 0.1 M phosphate buffer at pH 7.4 at ambient lab temperature and allowed to equilibrate for at least 1 day. The data in our databases refers to the logarithm of solubility expressed in mol $L^{-1}$ (M). Protein binding is measured using the equilibrium dialysis technique. The compound is added to 10% plasma giving a concentration of 20 $\mu$M and dialyzed with isotonic buffer for 18 h at 37 °C. The plasma and buffer solutions are analyzed using generic LC/MS, and the first apparent binding constant for the compound derived. The binding constant is then used to determine the $\%_{free\ drug}$ in 100% plasma. Rat serum protein binding data correspond to the apparent binding equilibrium constant extrapolated to 100% plasma. LogD data were measured at pH 7.4 using the classical shake flask method. After filtration of the data, 34 484 compounds with a measured solubility, 27 828 with a measured logD, and 12 651 with a measured protein binding were retrieved. The number of compounds for which at least one of these three properties was defined was 54 427. The calculated descriptors included ClogP (as implemented in the Daylight toolkit),[18] molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, van der Waals polar surface area, and the number of violations of the Lipinski's rule-of-5. These were calculated for the 54 427 compounds retrieved. In Table 1, the average properties of the compounds used in the database are summarized.

The average values are not surprising considering the nature of the dataset. As expected for druglike molecules, which are mostly targeted toward oral administration, the average values respect Lipinski's rule-of-5. This is shown by the zero median value of the number of violations to that rule. The average solubility is 24.6 $\mu$M. The values corresponding to the average plus standard deviation and minus standard deviations are 1.8 and 416.9 $\mu$M, respectively. The mean solubility is low but reflects the presence of compounds

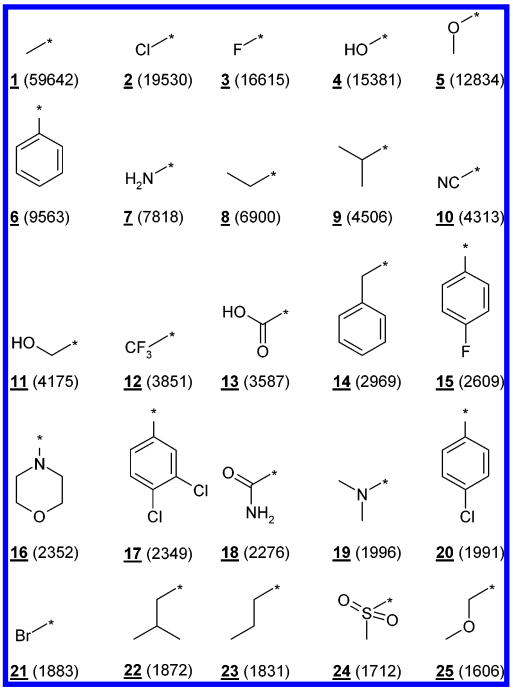



**Figure 2.** Twenty-five most represented side-chains in our database. The number below each side-chain corresponds to its ranking, and the number in parentheses refers to its frequency of occurrence.

synthesized during the optimization process and for which less-stringent criteria are applied. The average of $\%_{free}$ derived from the log $K1_{app}$ is 2.7%, in good agreement with druglike compounds. Protein binding is not considered until late in the drug discovery process. This is reflected by the values corresponding to the average plus standard deviation and minus standard deviations, 0.4 and 17.3%, respectively. These values are too low and too high for a drug, respectively. All these data were stored in a MySQL database. New data is continuously added to corporate databases; hence, the figures presented here only correspond to a snapshot. An update of our system would report different figures.

**Side-Chain Dataset.** In contrast to other studies,[19] a list of side-chains was predetermined. Hence, when the user submits a molecule, only side-chains that are present in that list can be detected. This limitation was essentially taken for computational tractability reasons. An on-the-fly determination of historical replacements would have been too computationally intensive and therefore not amenable to a web-interface. Hard coding substitutions imposed a finite number of side-chains. Starting from the AstraZeneca corporate collection and using software developed in-house and based on the methodology described by Lewell,[20] we defined a list of side-chains. Side-chain fragments that broke cycles and those containing more than 40 characters in the SMILES notation were removed from the list. A total of 9038 side-chain fragments were determined this way. The frequency of occurrence of these fragments in the 54 427 compounds dataset has been determined. The most represented functional groups are described in Figure 2, together with their ranking and an indication of their frequency of occurrence.

This ranking does not take into consideration the type of atom to which the fragment was attached. They are quite similar to those determined by Ertl[11] who processed a

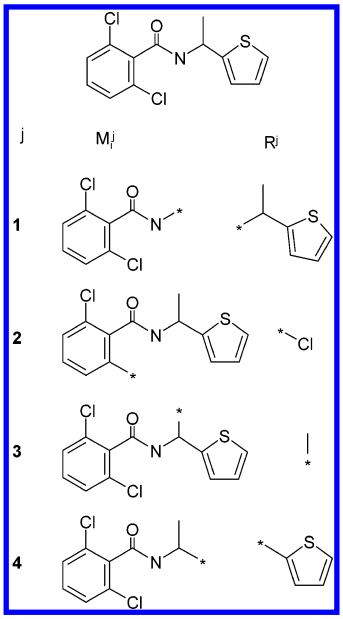



**Figure 3.** Nonexhaustive example of decompositions of one molecule. The example molecule is illustrated on the top of the figure. On the side-chains $R^j$, the asterisk indicates the atom to which the side-chain is attached. On the remaining of the molecule $M_i^j$, the asterisk indicates the point of attachment of the side-chain.

database of ~3 million compounds and identified 850 000 unique substituents. In the 25 most commonly occurring residues of our study, only three are not present in the most-common substituents identified in Ertl's study. Among them 3,4-dichlorophenyl, ranked 17, has an isomer in Ertl's work: 2,4-dichlorophenyl. Bemis[9] did a similar study on the comprehensive medicinal chemistry database and identified ~19 000 side-chains. Their definition of framework[21] excluded side-chains containing cycles. If we exclude side-chains containing cycles from our ranking, we are left with 19 functional groups. Among these 19 residues, 17 are present in the ranking of Bemis. This indicates that the identified substituents are effectively the most common substituents. In contrast to Ertl and Lewell's studies, where ~64% of substituents were present only in a single molecule, in our work just 40% of side-chains are singletons. This smaller figure can be partly explained by the nature of the compound dataset used. Whereas public databases of commercially available compounds only contain molecules that reached the market or have been published, our in-house database contains series of compounds and not just the final molecule going into development. The diversity within these series is legitimately lower, hence explaining the lower rate of singletons. Out of the 9038 side-chains identified, 79 fragments are truly common, that is, present in more than 1% of the molecules in our database, while 684 are present in more than 0.1%. These figures are a bit larger than the ones reported in the Ertl study, reflecting the somewhat smaller diversity of our dataset.

## SUBSTITUTION

**Compound Decomposition.** To define a substitution or a replacement in a molecule, we first need to define a fixed part and a varying part or side-chain. For each molecule, a pattern search is performed to identify side-chains from a predefined list. If one side-chain $R^j$ is spotted, then the molecule can be split in two parts, the side-chain $R^j$ and the remainder of the molecule $M_i^j$. Molecule $M_i$ can therefore be written as $M_i^j + R^j$. If more than one side-chain is identified, several combinations exist. Typical decompositions for a given molecule are illustrated in Figure 3.

**Table 2.** Distribution of Anchoring Atoms

| anchoring atom | % |
|---|---|
| C | 37.1 |
| N | 12.5 |
| C aromatic | 37.2 |
| N aromatic | 2.7 |
| O | 7.8 |
| P | < 0.1 |
| S | 2.7 |
| Si | <0.1 |

**Table 3.** Percentage of Pairs of Side-Chains Represented by *n* Pairs of Molecules

| *n* | SOL74 (%) | logD (%) | SPB (%) | calcd (%) |
|---|---|---|---|---|
| 1 | 70.5 | 85.3 | 74.0 | 76.5 |
| 2 | 16.6 | 10.2 | 16.4 | 13.7 |
| 3 | 5.2 | 2.1 | 4.8 | 3.6 |
| 4 | 3.2 | 0.9 | 2.3 | 2.1 |
| 5 | 1.1 | 0.4 | 0.7 | 0.8 |
| 6 | 1.0 | 0.3 | 0.8 | 0.7 |
| 7 | 0.5 | 0.2 | 0.2 | 0.3 |
| 8 | 0.4 | 0.1 | 0.2 | 0.3 |
| ≥ 9 | 1.4 | 0.5 | 0.7 | 1.0 |

This decomposition was performed for all molecules present in the compound dataset. Every time a combination is identified, the type of atom to which the side-chain fragment is attached is recorded. Anchoring atoms are defined using the SMILES notation,[17] that is, C for an sp3 carbon atom, c for an aromatic carbon, N for a nonaromatic nitrogen atom, n for an aromatic nitrogen, O for an oxygen atom. With the complete compound dataset, 386 757 combinations were identified. More than 70% of functional groups are anchored to carbon atoms, with an equivalent contribution from aliphatic and aromatic carbon atoms. The distribution of anchoring atoms is summarized in Table 2.

**Identification of Substitutions and Calculation of Average Change.** The list of combinations determined in the decomposition is then analyzed to identify substitutions. By substitution between two molecules $M_1$ and $M_2$, we mean a change of side-chain where the remainder of both molecules are identical. If $M_1 = M_1^a + R^a$, $M_2 = M_2^b + R^b$, and $M_1^a = M_2^b$, then a substitution of $R^a$ by $R^b$ has been observed between $M_1$ and $M_2$. If we then define $X(M_1)$ to be a property of molecule $M_1$, then the change of that property associated with substitution of $R^a$ by $R^b$ for that particular pair of molecules is

$$\delta_X^{1,2}(R^a,R^b) = X(M_1) - X(M_2)$$

The average change of property $X$ associated with substitution of $R^a$ by $R^b$ is

$$\mu_X(R^a,R^b) = \langle \delta_X^{i,j}(R^a,R^b) \rangle_{(i,j)}$$

where $\langle\ \rangle_{(i,j)}$ denotes an average over all pairs of molecules where substitution of $R^a$ by $R^b$ has been observed. Identical side-chain replacements might be observed where the side-chains were attached to different atoms to avoid the sug-

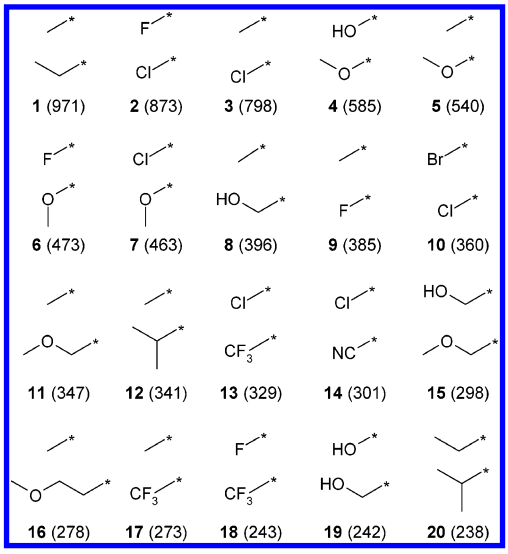



**Figure 4.** Twenty most represented side-chain replacements observed in the entire database. The number below each side-chain corresponds to its ranking, and the number in parentheses refers to its frequency of occurrence.

gestion of unphysical molecules in the remainder of the process, anchoring atoms are recorded and averages are performed individually for each class of atom. The average for pair ($R^a$,$R^b$) attached to an $sp^3$ carbon can therefore be defined as

$$\mu_X(R^a,R^b,C) = \langle \delta_X^{\,i,j}(R^a,R^b,C) \rangle_{(i,j)}$$

Similarly, a standard deviation $\sigma_X(R^a,R^b,C)$ and a number of observations $N_X(R^a,R^b,C)$ can be defined. Among the 386 757 combinations, 733 445 side-chain substitutions were identified, of which 336 817 were relative to solubility, 432 805 were relative to logD, and 198 342 were relative to serum protein binding. Fields in the table included the two side-chains occurring in the replacement $R^a$ and $R^b$, the anchoring atom $\chi$, and for each property, the average change historically observed $\mu_X(R^a,R^b,\chi)$, the standard deviation $\sigma_X(R^a,R^b,\chi)$, and the number of observations $N_X(R^a,R^b,\chi)$. As described in Table 3, when solubility data are considered, 70.5% of observed substitutions have been observed only once, that is, they correspond to a unique pair of molecules. Only 3.3% of observed substitutions in the solubility dataset have been observed more than five times. This ratio drops to 1.1% when looking at the logD dataset and 2.3% when all compounds are considered. The ratio of singletons is high, but this is the price to pay to be able to compare shifts in experimental properties.

The 20 most frequently observed side-chain replacements are detailed in Figure 4 for the entire database. Their rankings are indicated accompanied by an indication of their frequency of occurrence inside parentheses.

Most of these replacements are classical replacements observed in medicinal chemistry. Nevertheless, it is quite interesting to note that different results are obtained compared to Sheridan[13] who analyzed replacements of chemical groups in pairs of molecules from the MDDR[14] with the same biological activity. Sheridan's study was not limited to substituents; comparison was therefore limited to fragments that correspond to our substituent definition. While some replacements are common to both studies, like methyl to chlorine, methyl to methoxy or fluorine to methoxy, most replacements observed in Sheridan's study do not appear in the 20 most frequently occurring replacements of our study. This reflects the fact that only bioisosteric replacements were considered, whereas all replacements regardless of of their biological activity were considered in this work. These considerations biased substitutions toward functional groups of similar size and physicochemical properties in the Sheridan study. We also examined the most frequent replacements with a dataset restricted to compounds possessing solubility data (data not presented); their examination did not reveal any significant difference from the global ranking. We were expecting this classification to highlight substitutions containing solubilizing groups; this was not the case. This would have been the case if chemical series in that dataset were synthesized to improve solubility. While this is true for some series that have issues with solubility, this is obviously not the case for all of them.

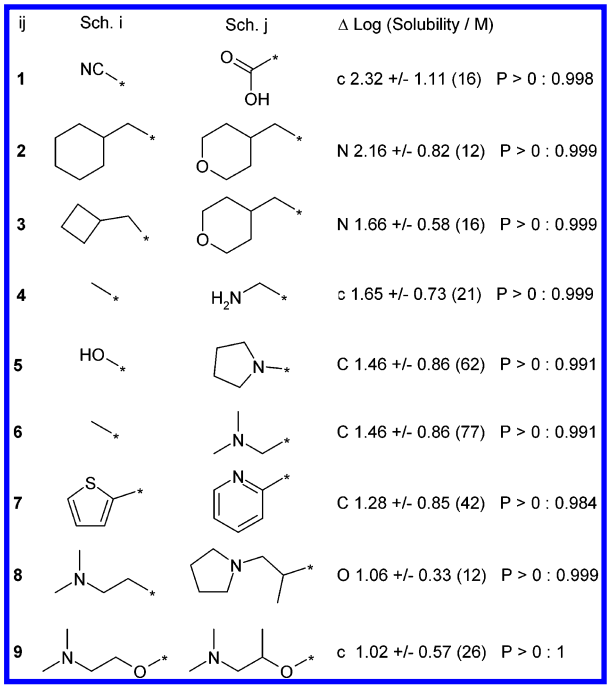



**Figure 5.** Selected side-chain replacements which have been observed to increase solubility. This selection is restricted to pairs for which at least 10 pairs of molecules have been used in the average and the confidence interval is at least 68%, that is, $\mu - \sigma > 0$, where $\mu$ corresponds to the mean of the selected pair and $\sigma$ to its standard deviation. The anchoring atom, average, and standard deviation, together with the number of observations, inside parentheses, are indicated on the right of each pair. The one-tailed probability confidence interval $P$ ($x > 0$) is also indicated. Calculated using $P(\mu - n\sigma < x < \mu + n\sigma) = \mathrm{erf}(n/\sqrt{2})$, where $\mathrm{erf}(x)$ is the so-called error function, $n$ was defined as $\mu - n\sigma = 0$, that is, the lower confidence limit is zero.

A selection of substitutions that we have found to have been historically associated with an increase in aqueous solubility is displayed in Figure 5. This set only contains side-chain replacements that have been observed in at least 10 pairs of molecules, and for which, the average is significantly greater than zero. The probability confidence interval, $P(x > 0)$ is indicated and is at least greater than 98%. Not surprisingly, most of these fragment pairs involve replacements of a functional group by a more polar entity at pH 7.4. Typical replacements involve substitution of a neutral species at physiological pH by a tertiary amine (**4**, **5**, **6**), or

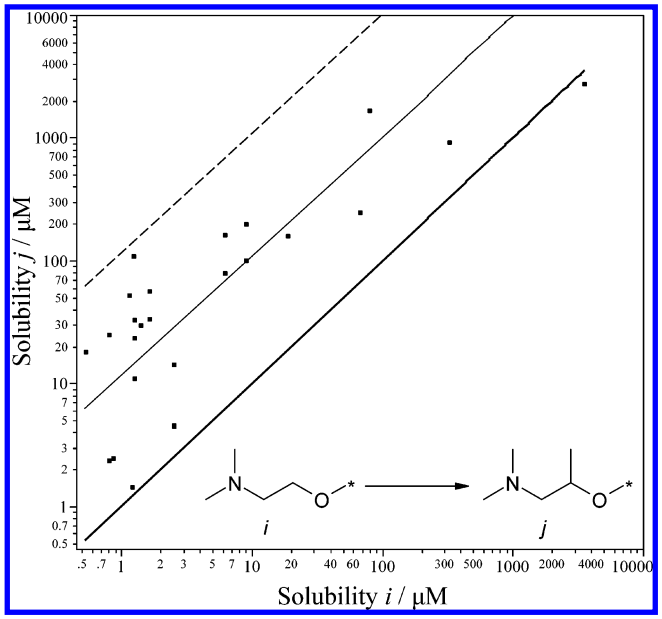



**Figure 6.** Plot of solubility vs solubility. Each point corresponds to a pair of molecules where the substitution labeled **9** in Figure 5 has been observed. The bold line is the first diagonal; the thin line corresponds to 10-fold increase in solubility, and the dashed line refers to a 100-fold increase.

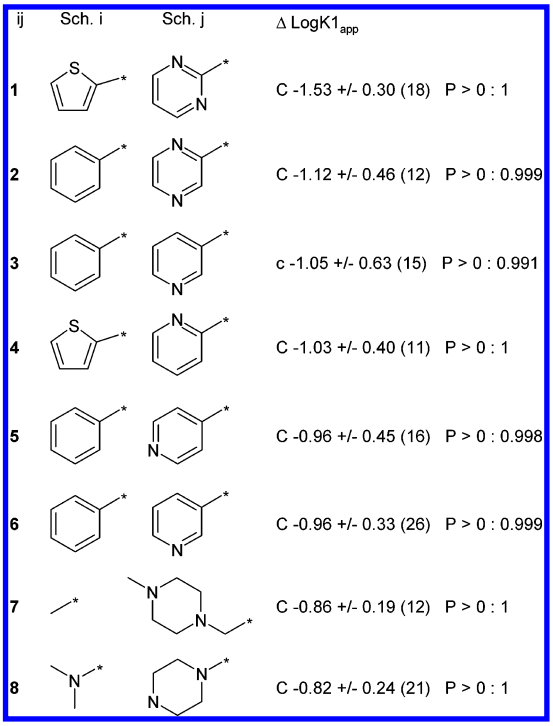



**Figure 7.** Selected side-chain replacements which have been observed to decrease log $K1_{app}$ (or increase $\%_{free}$). This selection is restricted to pairs for which at least 10 pairs of molecules have been used in the average and the confidence interval is at least 68%, that is, $\mu - \sigma > 0$, where $\mu$ corresponds to the mean of the selected pair and $\sigma$ to its standard deviation. The anchoring atom, average, and standard deviation, together with the number of observations, inside parentheses, are indicated on the right of each pair. The one-tailed probability confidence interval $P$ ($x > 0$) is also indicated. Calculated using $P(\mu - n\sigma < x < \mu + n\sigma) = \text{erf}(n/\sqrt{2})$, where erf(x) is the so-called error function, $n$ was defined such as $\mu - n\sigma = 0$, that is, the lower confidence limit is zero.

replacement by a more basic entity (**8**). Replacements of a neutral species by an acid (**1**) are present but less frequent than basic-group substitutions. Replacement **9** is particularly interesting because the difference between the two functional groups is a single methyl group. A simple methyl, sometimes referred to as magic methyl,[22] can disrupt crystal lattices or break hydration spheres. While this addition renders the amine function more basic (in-house measurements on those side-chains attached to identical aromatic ring indicate a shift of p$K_a$ from 7.8 to 7.9), the change of basicity cannot on its own explain the observed increase of solubility of one log scale unit. Indeed, the contribution of p$K_a$ on the total solubility is $\log(1 + 10^{pKa-pH})$. [The total solubility $S$ of a monobase at any pH is governed by the equation $S = [B] + [BH^+] = S_0(1 + [H^+]/K_a)$ or $\log S = \log S_0 + \log(1 + 10^{pKa-pH})$, where $S_0$ is the intrinsic solubility.] The influence on the total solubility that can be attributed to the p$K_a$ at pH = 7.4 is therefore $\log(1 + 10^{pKa-7.4})$.

Hence for p$K_a$ = 7.8 and 7.9 at physiological pH, p$K_a$ contributions to log $S$ are 0.545 and 0.619, respectively. The increase in p$K_a$ can therefore be associated with an increase of log $S$ of 0.074, which does not account for the observed average increase. Most of the observed increase in solubility can be accounted for by the intrinsic contribution. Illustrated in Figure 6 are the pairs of molecules involved in substitution **9**. Nearly half of the compound pairs exhibit a solubility increase of 0−1 orders of magnitude, with the other half being associated with 1−2 orders of magnitude increase. There is only one pair of molecules associated with a decrease in solubility going from 3540 to 2710 $\mu$M (the point located on the top right corner of the figure). These two numbers are actually the same within error.

Selected side-chain substitutions that have been associated with an increase in $\%_{freedrug}$ or decrease in log $K1_{app}$ are illustrated in Figure 7. Typical replacements increasing $\%_{freedrug}$ involve substitution of an aromatic substituent by a more polar aromatic ring, thiophene to pyrimidine (**1**) or pyridine (**4**), benzene to pyrazine (**2**) or pyridine (**3**, **5**, **6**). The position of the nitrogen atom on the pyridine ring (para (**5**) or meta (**6**)) has little influence on the observed values, the corresponding averages not being statistically different. The last substitution (**8**) involves the replacement of a basic group by a less basic group, in that case the more hydrophilic nature of the *n*-methyl piperazine governs the increase in $\%_{freedrug}$ rather than the basicity of the amine.

## WEB INTERFACE

Numerous historically observed replacements in druglike molecules, together with their associated change in physicochemical properties have been identified. Even if they are not all relevant or useful, they still represent a wealth of data that falls far beyond the reach of one medicinal chemist's knowledge. This information combined with chemist's knowledge, intuition, and chemical sense would represent a fantastic tool for lead optimization. At this stage of the drug discovery process, subtle modifications are brought to a compound in order to bring it to the status of drug candidate. An increase in biological activity is always welcome; nevertheless activity is supposedly already achieved at that stage. Modifications are therefore mainly brought so that physicochemical properties and PK/PD behavior fits with the targeted drug profile. The scaffold of the lead is not theoretically modified, changes mostly include replacement of side-chains. A crystal structure of the target protein might

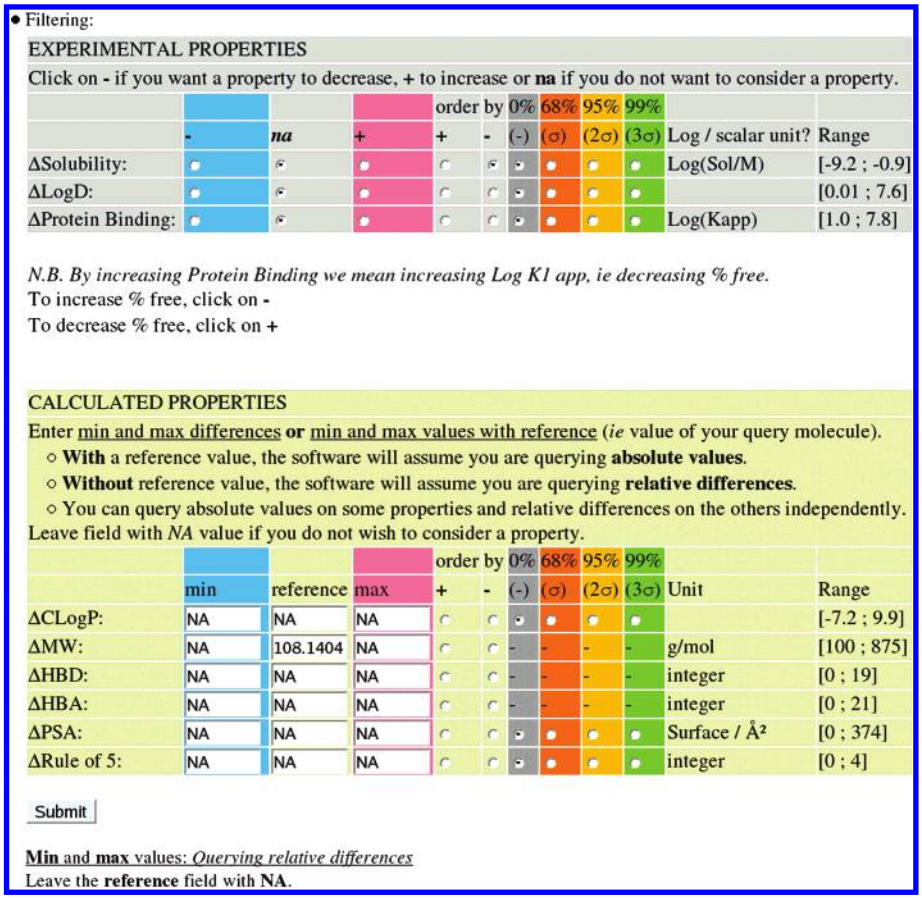



**Figure 8.** Filters available during the substituent selection process.

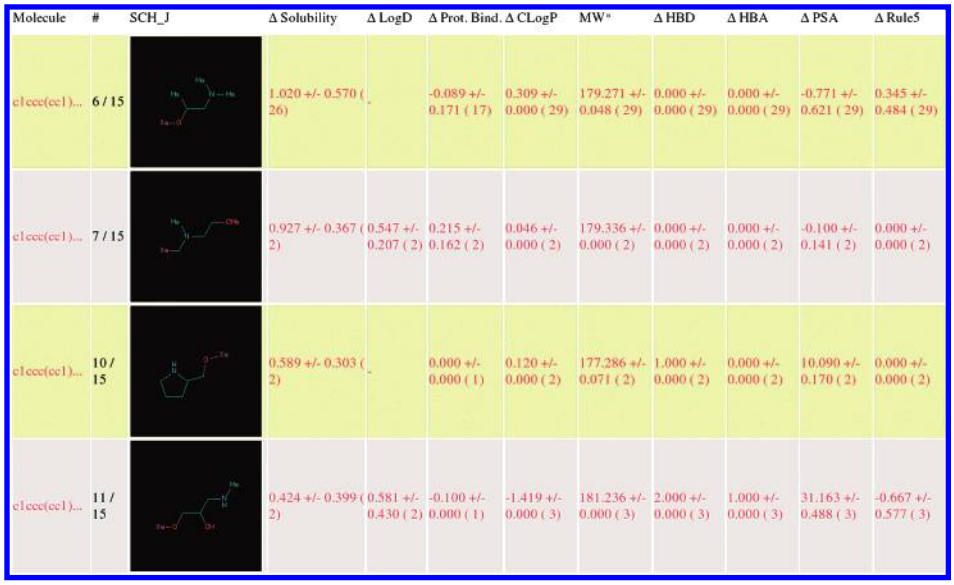
even be known, therefore restricting the modifications that can be made to the compound. Some side-chains interact specifically with the target and should therefore not be modified. Replacements of other groups are required to modulate the compound properties. Whereas in the concept of bioisosterism, replacement of a functional group by another group possessing similar size and physicochemical properties is desired, in the present work, we try to identify chemical replacements that increase or decrease certain physicochemical properties. We therefore developed a tool that enables a chemist to identify replacements of side-chains that have occurred in the past and which have been associated with an increase or decrease in any given property (filters). The user feeds in a compound, and the software returns a list of side-chain pairs fulfilling the user's filters together with the new molecules built from the identified side-chains. We implemented this procedure as a web interface because it represents an ideal environment for chemical information processing.[23] The process, which is described in Figure 1, can be decomposed into five steps.

For step 1, the user feeds in the molecule to be optimized: $M_1$. This is either done through the JME molecular editor or by pasting SMILES into the entry box. In both cases, the molecule is passed to the server in SMILES format.

During steps 2 and 3, the software determines side-chains present in the molecule and determines identified combinations, $M_1 = M_1{}^a + R^a = M_1{}^b + R^b = M_1{}^c + R^c = ...$, one for each side-chain identified. The different side-chains that can be detected are predetermined. They are taken from a dataset comprising 9038 functional groups that have been described previously. The list of identified side-chains is returned to the user.

For step 4, the user selects the identified side-chains he wishes to replace and enters filters (if required) on solubility, logD, protein binding, number of hydrogen bond donor/acceptor, molecular weight, logP, polar surface area, and number of violations to Lipinski's rule-of-5. The filter window is illustrated in Figure 8. Two kinds of filters can be applied, filters on experimental properties and filters on calculated properties. For experimental properties, the user specifies whether he wants a particular property to increase or decrease (gray box in Figure 8). This option is the simpler to comprehend for the user and corresponds to most requests. When one of these options is selected, the system only returns substitutions that have been associated with an increase/decrease in the property selected.

On a second level, filters applied to calculated properties (light-green box in Figure 8) are a bit more sophisticated and allow the user to apply ranges, either on differences of properties or on the absolute values of the suggested molecules. If the user does not provide a reference value, the system will assume that query is performed on the difference. Queries on differences and absolute values can be performed independently for each parameter. For example, one might request substitutions that have been associated with an increase in solubility, for which ClogP differences are between 1 and 3 and where the target molecule molecular weight is lower than 500 g $mol^{-1}$. The molecular weight of the input molecule is calculated from the full atom representation. The latter is generated using smi2mdl,[24] a software that allows the transformation of a SMILES into a mol2 file.



| Molecule | # | SCH_J | Δ Solubility | Δ LogD | Δ Prot. Bind. | Δ CLogP | MW* | Δ HBD | Δ HBA | Δ PSA | Δ Rule5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| c1ccc(cc1)... | 6 / 15 | | 1.020 +/- 0.570 ( 26) | | -0.089 +/- 0.171 ( 17) | 0.309 +/- 0.000 ( 29) | 179.271 +/- 0.048 ( 29) | 0.000 +/- 0.000 ( 29) | 0.000 +/- 0.000 ( 29) | -0.771 +/- 0.621 ( 29) | 0.345 +/- 0.484 ( 29) |
| c1ccc(cc1)... | 7 / 15 | | 0.927 +/- 0.367 ( 2) | 0.547 +/- 0.207 ( 2) | 0.215 +/- 0.162 ( 2) | 0.046 +/- 0.000 ( 2) | 179.336 +/- 0.000 ( 2) | 0.000 +/- 0.000 ( 2) | 0.000 +/- 0.000 ( 2) | -0.100 +/- 0.141 ( 2) | 0.000 +/- 0.000 ( 2) |
| c1ccc(cc1)... | 10 / 15 | | 0.589 +/- 0.303 ( 2) | | 0.000 +/- 0.000 ( 1) | 0.120 +/- 0.000 ( 2) | 177.286 +/- 0.071 ( 2) | 1.000 +/- 0.000 ( 2) | 0.000 +/- 0.000 ( 2) | 10.090 +/- 0.170 ( 2) | 0.000 +/- 0.000 ( 2) |
| c1ccc(cc1)... | 11 / 15 | | 0.424 +/- 0.399 ( 2) | 0.581 +/- 0.430 ( 2) | -0.100 +/- 0.000 ( 1) | -1.419 +/- 0.000 ( 3) | 181.236 +/- 0.000 ( 3) | 2.000 +/- 0.000 ( 3) | 1.000 +/- 0.000 ( 3) | 31.163 +/- 0.488 ( 3) | -0.667 +/- 0.577 ( 3) |

**Figure 9.** Selected results of a query for the first substituent of pair **9** from Figure 5, increasing solubility with a 68% confidence interval.

Calculation of other parameters has not been implemented yet and will be the goal of future development of the system. Users therefore have to enter ClogP, PSA, HBD, HBA, and number of violations to the rule-of-5 if they want to filter the resulting compounds on these properties.

Identical substitutions can be associated with an increase or decrease in any given experimental property. Solubility, for example, is delicate to understand, a simple modification can have large effects in some cases and opposite effects in others. To balance that effect, we introduced the ability for the user to filter results using confidence intervals. A confidence interval[25] is an interval in which a measurement or trial falls corresponding to a given probability. Usually, the confidence interval of interest is symmetrically placed around the mean. For a normal distribution, the probability that a measurement falls within $n$ standard deviations ($n\sigma$) of the mean (i.e., within the interval $[\mu - n\sigma, \mu + n\sigma]$) is $P(\mu - n\sigma < x < \mu + n\sigma) = 0.68$ for $n = 1$, 0.95 for $n = 2$, and 0.99 for $n = 3$. For a given side-chain replacement with a mean $\mu$ and standard deviation $\sigma$ and if the hypothesis is that the property is increased by that substitution (i.e., the shift is effectively greater than zero), then the observed shift can be said to be significant at the 68% confidence level if $\mu - \sigma > 0$. In that case, the probability that the measurement is greater than zero would actually be 0.84 (one-tailed) rather than 0.68 (two-tailed). Internally, these filters are applied using simple SQL queries. Confidence intervals filters can be applied for all properties, experimental and calculated, with the exception of molecular weight and the number of hydrogen bond donors and acceptors. Indeed, these differences do not have standard deviation.

In step 5, following the SQL query, the server returns a list of historically observed replacements of the selected side-chains fulfilling the criteria set out by the filters, as well as the new molecules built from the found side-chains. This step is illustrated in Figure 9.

Finally an optional step might be performed; for each pair of side-chains, the user can query the database to retrieve the list of molecule pairs where the replacement has been observed. Instead of storing the list of compounds used for each pair, the list is generated on the fly using an SQL query on both the table containing the combinations and on the table containing the experimental and calculated properties.

All these steps are performed on a single Linux computer, which holds both the Apache web server and MySQL databases. Processing of web requests is based on cgi scripts written in PERL, which call various C and PERL executables to process the user's compounds. Most of the scripts and software used have been developed in-house at AstraZeneca. External packages used include smi2mol2 and mol2gif by OpenEye[24] to generate picture of molecules and to convert molecules into MDL mol2 format. Molecular entry is performed with the JME molecular editor.[26]

## DISCUSSION

We presented the development of a substitution database together with a web-interface allowing AstraZeneca medicinal chemists to query it on the corporate intranet. This approach presents several advantages. The proposed method is computationally efficient when compared to several other methods.[27] During the database setup, calculations of computationally intensive parts are performed on a *per* molecule basis rather than on a *per* pair of molecule basis. The setup process therefore scales as $N$ rather than as $N^2$, where N is the number of compounds initially analyzed. Indeed the most intensive part of the process concerns the step where each compound is decomposed into (substituent/remaining part of the molecule) combinations. The comparison of the combinations and identification of substitutions is just a text-based match and does not constitute the computer-intensive part of the setup. The nature of the interface and of the database ensures fast queries and access across all platforms.

One of the key features of the proposed methodology is the fact that it is essentially built on real data rather than calculated properties. While this approach does not allow us to cover the amount of chemical space covered by methods based on calculated properties, the effects which are reported are real and carry more weight. The database can be updated automatically, therefore reflecting new functional groups brought to the collection. This methodology might also easily be applied to other experimental properties.

After they have performed a search and identified possible chemical replacements, users have the added ability to retrieve the list of compound pairs where a particular substitution was observed. This represent a SAR tool of great value to medicinal chemists, as well as a knowledge management tool in the sense that it represents a quick way of accessing information relating to particular substitutions filtered in the direction desired by the chemist.

One of the objectives will now be to apply the described methodology to other external databases such as MDDR. This would allow users to explore new parts of the chemical space. Another application to scaffold hoping as described by Barker[28] is also considered.

## ACKNOWLEDGMENT

D.Y.H. would like to thank David Cosgrove for precious help on software-related issues, as well as Peter Ertl for providing the JME molecular editor. The authors would like to thank David Cosgrove and Jens Sadowski for providing AutoRSgroper, Splitter, and Scaffsplit software. The authors would also like to thank Paula Kitts, Andrew Leach, and Han van de Waterbeemd for proofreading this manuscript.

## REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening—An overview. *Drug Discovery Today* **1998**, *3* (4), 160−178.
(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46* (1−3), 3−26.
(3) Reichardt, C. In *Solvents and Solvent Effects in Organic Chemistry*, 2nd ed.; VCH Verlagsgesellschaft: Weinheim, Germany, 1988; Chapter 7, pp 27−35.
(4) Lewis, R. A. A general method for exploiting QSAR models in lead optimization. *J. Med. Chem.* **2005**, *48* (5), 1638−1648.
(5) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.* **2006**, *46* (1), 180−192.
(6) Bruneau, P. Search for predictive generic model of aqueous solubility using Bayesian neural nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1605−1616.
(7) Gleeson, M. P.; Waters, N. J.; Paine, S. W.; Davis, A. M. In silico human and rat Vss quantitative structure-activity relationship models. *J. Med. Chem.* **2006**, *49* (6), 1953−1963.
(8) Zuegge, J.; Schneider, G.; Coassolo, P.; Lave, T. Prediction of hepatic metabolic clearance: comparison and assessment of prediction models. *Clin. Pharmacokinet.* **2001**, *40* (7), 553−563.
(9) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42* (25), 5095−5099.
(10) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511−522.
(11) Ertl, P. Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of druglike bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 374−380.
(12) Patani, G. A.; LaVoie, E. J. Bioisosterism: A rational approach in drug design. *Chem. Rev.* **1996**, *96* (8), 3147−3176.
(13) Sheridan, R. P. The most common chemical replacements in druglike compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 103−108.
(14) Molecular Design. Drug Data Report. http://www.mdl.com/products/knowledge/drug_data_report/ (accessed March 26, 2007).
(15) Advanced Chemistry Development. Structure Design Suite. http://www.acdlabs.com/products/phys_chem_lab/structuredesign/ (accessed March 26, 2007).
(16) (a) Kassam, K.; Sasaki, R.; Hachey, M. R. J. Reducing the synthetic burdens of lead structure optimization: a novel software-aided approach; Presented at the DDT Conference, Boston, MA, 2005. (b) ACD labs. http://www.acdlabs.co.uk/download/publ/2005/ddt05_sds.pdf (accessed March 26, 2007).
(17) Weininger, D. SMILES, A chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(18) Daylight Toolkit. http://www.daylight.com/products/toolkit.html (accessed March 26, 2007).
(19) Cosgrove, D. A.; Willett, P. SLASH: A program for analysing the functional groups in molecules. *J. Mol. Graphics Modell.* **1998**, *16* (1), 19−32.
(20) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; McLay, I. M.; Bradshaw, J. Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46* (15), 3257−3274.
(21) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(22) Smith, D. A.; van de Waterbeemd, H. Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* **1999**, *3* (4), 373−378.
(23) Ertl, P.; Muhlbacher, J.; Rohde, B.; Selzer, P. Web-based cheminformatics and molecular property prediction tools supporting drug design and development at Novartis. *SAR QSAR Environ. Res.* **2003**, *14* (5−6), 321−328.
(24) OpenEye Scientific Software OEChem. http://www.eyesopen.com (accessed March 26, 2007).
(25) Kenney, J. F.; Keeping, E. S. Confidence limits for the binomial parameter and Confidence interval charts. In *Mathematics of Statistics*, Part 1, 3rd ed.; Van Nostrand: Princeton, NJ, 1962; Chapter 11.4, 11.5; pp 167−169.
(26) Ertl P. JME molecular editor applet allowing creation or editing of molecules. http://www.molinspiration.com/jme/ (accessed March 26, 2007).
(27) Sheridan, R. P.; Miller, M. D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Model.* **1998**, *38* (5), 915−924.
(28) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold hopping using clique detection applied to reduced graphs. *J. Chem. Inf. Model.* **2006**, *46* (2), 503−511.

CI600395U