

Successful identification of key chemical structure modifications that lead to improved ADME profiles

Lourdes Cucurull-Sanchez

Received: 22 October 2009 / Accepted: 26 April 2010 / Published online: 9 May 2010
© Springer Science+Business Media B.V. 2010

Abstract The results of a new method developed to identify well defined structural transformations that are key to improve a certain ADME profile are presented in this work. In particular Naïve Bayesian statistics and SciTegic FCFP_6 molecular fingerprints have been used to extract, from a dataset of 1,169 compounds with known in vitro UGT glucuronidation clearance, those changes in chemical structure that lead to a significant increase in this property. The effectiveness in achieving that goal of the thus found 55,987 transformations has been quantified and compared to classical medicinal chemistry transformations. The conclusion is that on average the new transformations found via in silico methods induce increases of UGT clearance by twofold, whilst the classical transformations are on average unable to alter that endpoint significantly in any direction. When both types of transformations are combined via substructural searches (SSS) the average twofold increase in glucuronidation is maintained. The implications of these findings for the drug design process are also discussed, in particular when compared to other methods previously described in the literature to address the question ‘Which compound do I make next?’

Keywords Pair wise comparisons · Reverse QSAR · UGT glucuronidation

Abbreviations

ADME	Absorption distribution metabolism and excretion
QSAR	Quantitative structure–activity relationships
UGT	Uridine 5′-diphospho-glucuronosyltransferase
ECFP	Extended connectivity finger-prints

Introduction

In silico QSAR methods in the pharmaceutical industry are typically used to obtain an estimate of some pharmacology or ADME endpoint purely from the molecular structure of a compound. The estimated value is generated by an in silico QSAR model, which could be viewed as the result of a training process consisting of the combination of three elements: (i) a set of measured values of the endpoint of interest for a number of compounds; (ii) a set of relevant molecular descriptors derived from the chemical structure of those compounds; and (iii) a statistical/machine learning method capable of establishing logical or mathematical relationships between (i) and (ii). When the model is properly trained and validated, a good estimate of the measured endpoint (i) can be derived from combining the molecular descriptors of any compound (ii) with the statistical rules learnt via the appropriate method (iii). At this point, the model can then be used by scientists to answer the question of ‘which endpoint value of Y would we obtain if we make a compound with X molecular features?’

Taking this view into account, an alternative application of the model training process could be defined, whereby a set of molecular descriptors (ii) would be obtained by combining a relevant type of endpoint values (i) with the statistical rules learnt via the appropriate method (iii). The

Electronic supplementary material The online version of this article (doi:10.1007/s10822-010-9361-5) contains supplementary material, which is available to authorized users.

L. Cucurull-Sanchez (✉)
Department of Pharmacokinetics, Dynamics and Metabolism,
Pfizer PGRD, Sandwich, Kent CT13 9NJ, UK
e-mail: Lourdes.Cucurull-Sanchez@Pfizer.com

outcome of such an algorithm would allow the scientists reverse the aforementioned question, i.e. ‘what sort of X molecular features do we need in our compound if we want to obtain an endpoint value of Y?’

The latter type of application of *in silico* QSAR methods (aka. ‘reverse QSAR’) has provided a number of rules of thumb designed to improve the quality of new drug entities (e.g. Lipinski’s rule-of-five [1], traffic light method for bioavailability [2], etc.). However, those analyses have almost exclusively been focused on physicochemical or ‘bulky’ molecular properties such as molecular weight, lipophilicity, atomic counts, etc. As Leach pointed out [3], usually the structural descriptor changes that describe how a compound should change to achieve the desired change in the endpoint property are not identified. In other words, those features that carry more specific chemical structure and connectivity information, such as molecular fingerprints, have been traditionally deemed too complex and relegated to an almost non-existent role in the establishment of QSAR rules.

This paper presents a successful case study where the use of molecular fingerprints in combination with *in silico* QSAR methods has provided a tool to identify key chemical structural modifications to improve the ADME profile of compounds designed to be administered via inhalation. In particular Naïve Bayesian statistics and SciTegic ECFP₆ molecular fingerprints have been used to extract, from a dataset with known *in vitro* UGT glucuronidation clearance, those changes in chemical structure that lead to a significant increase of UGT glucuronidation clearance. Such a set of structural modifications does not exist in the literature.

Materials and methods

The following steps were encoded in a Pipeline Pilot (v.7.5) [4] protocol, which takes ca. 2 h 30 min to run on a standard built of a Pipeline Pilot server on a Windows 2003 machine with 16 Gb of RAM and 16 (4 quad core) processors.

1. Data curation and preparation

The dataset used in this work consisted of 1,169 microsomal UGT glucuronidation clearance values measured on Pfizer proprietary compounds, according to procedures previously described in the literature [5]. The values ranged from <2 to >440 $\mu\text{l}/\text{min mg}$, with 536 measurements reported as censored data, i.e. corresponding to clearance values below or above the detection limits of the assay. The histogram in Fig. 1 shows the distribution of the dataset with and without censored values. It can be seen that the natural distribution of this endpoint is logged normal,

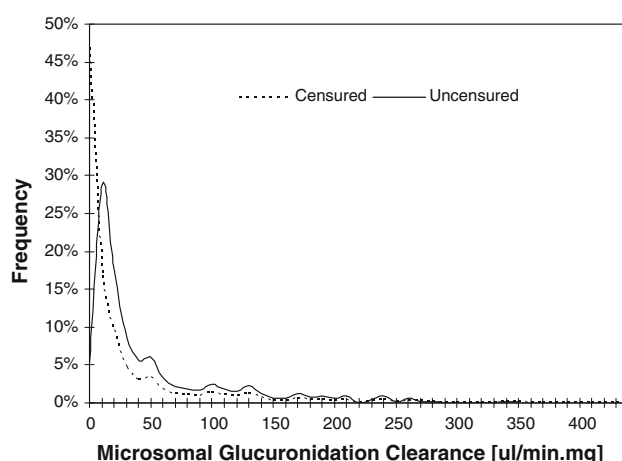


Fig. 1 Distribution of the endpoint values used in this study. The histogram shows how the microsomal UGT glucuronidation clearance values, ranging between <2 and >440 $\mu\text{l}/\text{min mg}$, follow a logged normal distribution. The dotted line represents the distribution of clearance values including those reported as ‘<’ or ‘>’ (aka. censored values) due to detection limits being reached, which represent 43% of the total dataset

consequently the geometric mean (26.2 $\mu\text{l}/\text{min mg}$) rather than the arithmetic mean was used as a criteria to split the data into high and low clearances. This resulted in the whole dataset being divided into two subsets of 870 low clearance (<26.2 $\mu\text{l}/\text{min mg}$) and 299 high clearance (≥ 26.2 $\mu\text{l}/\text{min mg}$) compounds.

2. Relevant feature extraction via *in silico* model building

Based on the entire curated data set, two *in silico* models were built with the ‘Learn Good Molecules’ component in Pipeline Pilot: one to learn the relevant features present in molecules with low UGT glucuronidation clearance (from now on referred to as the ‘Low Clearance’ model), and another to learn the features present in molecules with high UGT glucuronidation clearance (the ‘High Clearance’ model). This Pipeline Pilot component builds a Naïve Bayesian categorization model that distinguishes the active molecules for a certain target (typically referred to as the “good” data records) from the rest of the molecules available, known as the baseline set [6]. According to this, in the ‘Low Clearance’ model the “good” compounds were those contained in the low clearance subset, and the high clearance compounds were the baseline, whilst in the ‘High Clearance’ model these definitions were reversed. The type of molecular descriptors or features used in the model building processes was set to ECFP₆ fingerprints [7].

In addition to creating a new “model” component that can be employed prospectively to estimate the likelihood of a compound belonging to the “good” class, the ‘Learn Good Molecules’ component produces a list of features

whose presence in the “good” class is significantly larger than in the baseline compounds. Each feature in that list has an associated normalized probability value, which corresponds to the log value of the corrected probability of a compound being “good” when that feature is present [8, 9]. This means that a feature with high normalized probability will have more chances of leading to a “good” compound than a feature with low normalized probability. Those features with normalized probability scores close to zero are considered to be uninformative, so by default the Pipeline Pilot algorithm trims the feature list by removing those in the range $[-0.05, 0.05]$, in order to efficiently use the server disk space and memory. The ‘Low Clearance’ and ‘High Clearance’ modeling processes produced two ranked lists of 9824 and 3576 ECFP₆ fingerprint features respectively, each one summarizing the substructural features that are frequently and significantly observed in each one of the two compound subsets used as the “good” data records.

The fact that the features used to build the model are fingerprints, and therefore represent chemical fragments or substructures, enabled the extension of the feature trimming process beyond the default settings. So, to further enrich the list of features in those with the largest probability of leading to a “good” compound, only the “good” ECFP₆ fragments which scored positively according to their own models were selected for the analysis. In doing so, if a fingerprint happened to contain a sub-structure or a number of sub-structures that significantly weighted against a “good” behavior, its bayesian score (as opposed to its normalised probability score) would be negative and thus the fragment discarded from the final list of “good” features. This approach compensated for the “Naïve” assumption made during the development of the model, whereby features were presumed independent from each other [10]. The resulting ECFP₆ lists came down to 7,111 ‘low’ and 1,733 ‘high’ features.

At this stage, two lists of features had been collected. The first list was a compilation of the structural motifs that are most commonly found in compounds with ‘Low Clearance’. The second list summarised the chemical fragments that are significantly present in molecules with ‘High Clearance’. Being able to spot in a compound the particular sub-fragment(s) that makes it undesirable is of great importance. However being able to propose a substitute fragment that takes the compound towards the desired property space has an even larger significance for the drug designer. For this reason the next three steps were designed to find links between both lists of sub-structural features. The types of algorithms used involve sub-structural searches (SSS), exact searches, and combinations of both, as well as the incorporation of prior medicinal chemistry knowledge via previously defined chemical modifications.

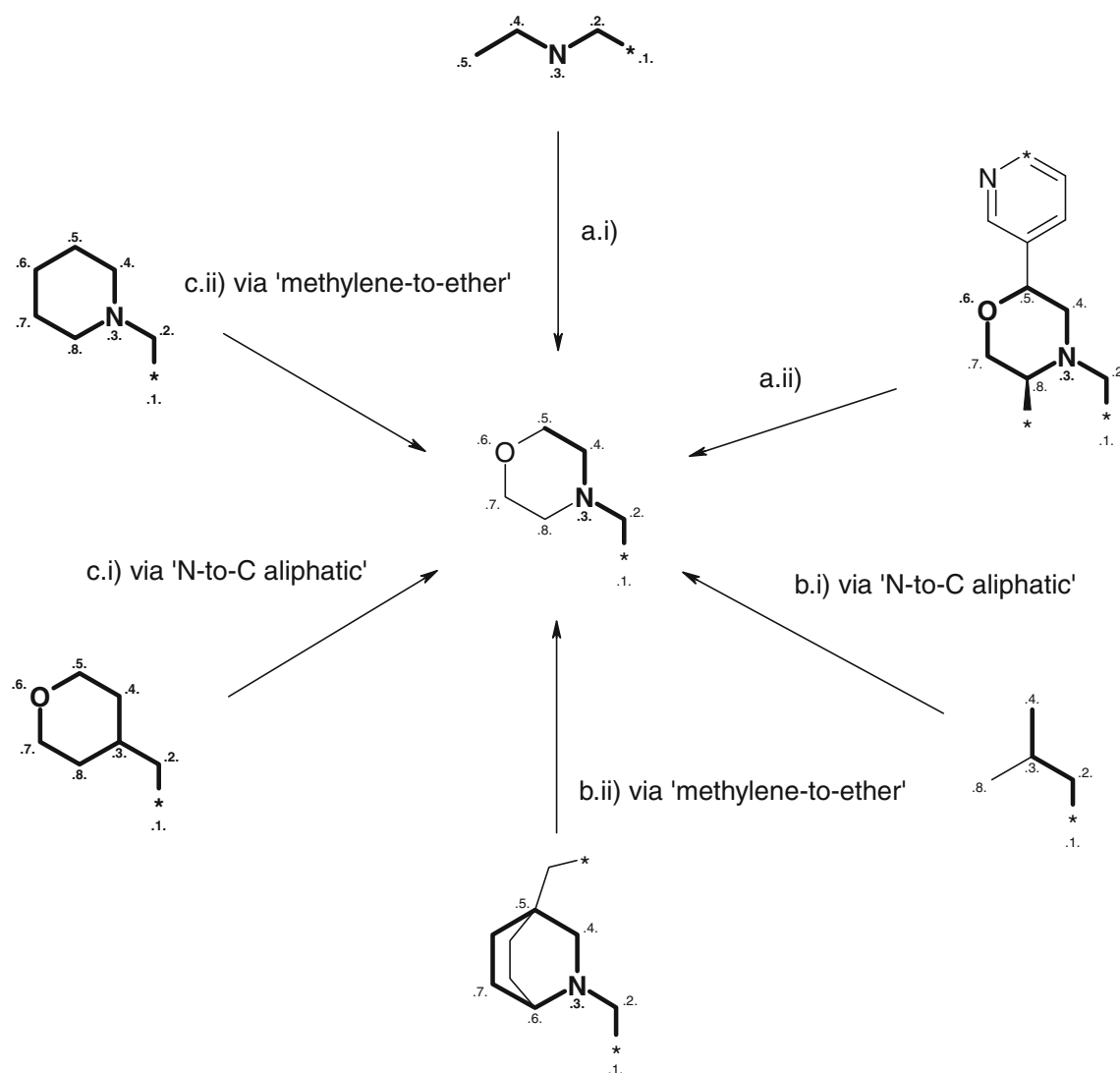
3. Feature sub-structural searches (‘Direct SSS Transformations’)

Since ECFP₆ fingerprints are a representation of molecular fragments, they can be used as queries or as targets of Sub-Structure Searching (SSS) algorithms. Based on this, the following two SSS processes were carried out:

- i. *Search of key structural modifications that increase the size of the molecule.* In this process, each ‘Low Clearance’ feature was used as a SSS query, to search for any mapping super-structure amongst the list of ‘High Clearance’ features (in SSS terms, the latter is often referred to as the ‘target’ set). Every successful mapping (in SSS terms known as ‘hit’) establishes a direct link between one ‘Low Clearance’ and one ‘High Clearance’ feature. This link represents the structural modification that can be applied to a low clearance feature in order to turn it into a high clearance feature, by way of *adding* some sort of molecular matter at certain points of the structure (an example of this can be seen in Scheme 1.a.i). So when a hit was found, the pair of linked ‘Low Clearance’ and ‘High Clearance’ features was stored in a database as a chemical reaction, where the query or ‘Low Clearance’ feature played the role of the reactant, and the matching target or ‘High Clearance’ feature played the role of the product.
- ii. *Search of structural modifications that reduce the size of the molecule.* Here the roles of each feature list in the search were exchanged, i.e. ‘Low Clearance’ features were now used as SSS targets and ‘High Clearance’ features as queries. In this case, the SSS hits represent a transition from a low to a high clearance chemical space via the *elimination* of some chemical matter from the molecular structure, as Scheme 1.a.ii illustrates. Like in the previous search, each hit was stored in a chemical reaction database as a reactant and a product pair.

4. Feature sub-structural searches combined with pre-defined medicinal chemistry transformations (‘Combined SSS Transformations’)

The process described in Sect. 3 was slightly modified in order to factor in prior knowledge about the feasibility and success of chemical modifications into the hit search. The new process combined the ECFP features found in step 2 with an in-house set of 440 transformations pre-defined by medicinal chemists. These transformations represent the replacements known to experienced medicinal chemists as ‘rules-of-thumb’ for drug design, like the Drug-GURU system described in the literature [11].



Scheme 1 Examples of key transformations found via three different mapping algorithms: (a) Feature Sub-Structural Search (SSS) to obtain 'Direct SSS Transformations' that increase (a.i) or decrease (a.ii) the size of the molecule; (b) Feature SSS combined with pre-defined medicinal chemistry transformations (e.g. 'methylene-to-ether' or 'N-to-C aliphatic') to obtain 'Combined SSS Transformations' that involve a traditional medicinal chemistry change either

prior to adding some molecular moiety (b.i) or after taking away some molecular sub-fragment (b.ii); and (c) Feature exact match search, combined with pre-defined medicinal chemistry transformations to obtain 'Combined Exact Transformations' that involve exclusively traditional chemical modifications (c.i or c.ii depending on whether the starting point for the modified queries were low or high clearance features)

Again the SSS took place in two different directions:

- i. *Search of key structural transformations with modified 'Low Clearance' features used as SSS queries.* The algorithm started with each 'Low Clearance' feature being modified via all the pre-defined drug-design transformations. This led to a number of structural fragments that were then used as queries for a SSS into the 'High Clearance' target space. When a hit was found, the 'High Clearance' feature was mapped to the original, unmodified 'Low Clearance' substructure and the pair stored as product and reactant, respectively.

The main difference between the structural changes found here and those found in Sect. 3.i., is that now reactant and product are linked by a common sub-structural core that has been modified in two steps, rather than just one: firstly, following standard practice amongst medicinal chemists, and secondly tuning the fragment to a high clearance chemical space via the addition of some extra chemical matter. This is illustrated by an example in Scheme 1.b.i, where the medicinal chemistry transformation is a replacement of a nitrogen atom in an aliphatic environment by a carbon atom.

- ii. *Search of key structural modifications with modified ‘High Clearance’ features used as SSS queries.* This search required the modification of all the ‘High Clearance’ features via the *reversed* pre-defined drug-design transformations. Then each modified ‘High Clearance’ feature was used as a SSS query against the ‘Low Clearance’ features target space. Like before, the process led to a matched pair of features that was stored as a chemical reaction, with a low clearance reactant and a high clearance product. In this type of searches, the ‘High Clearance’ substructure is the outcome of firstly removing some atoms from the ‘Low Clearance’ feature and secondly performing a standard medicinal chemistry transformation on the resulting fragment. This can be seen in the example in Scheme 1.b.ii, where the classical transformation applied is a change of a methylene group by an ether function.

5. Feature exact match search, combined with pre-defined medicinal chemistry transformations (‘Combined Exact Transformations’)

A faster alternative to step 4 is to perform an exact match search between modified and non-modified features, instead of a Sub-Structural Search. Although in principle the roles of query and target may seem equivalent for this type of search, the presence of explicit hydrogen atoms in some substructural features may cause mismatches that would otherwise correspond to structurally identical moieties. To avoid this problem and ensure that the search was exhaustive, the process was performed in two opposite directions:

- i. *Search of key structural transformations with modified ‘Low Clearance’ features used as exact match queries.* Similarly to step 4.i, the ‘Low Clearance’ features were modified via all the pre-defined drug-design transformations, and then used as queries against the ‘High Clearance’ features target space. However, because the aim of the current search was to find exact matches, a pair of substructures was considered a hit if and only if the all the atoms of the query mapped all the atoms in the target.
- ii. *Search of key structural modifications with modified ‘High Clearance’ features used as exact match queries.* The ‘High Clearance’ features were modified via all the pre-defined drug-design transformations, and then used as queries against the ‘Low Clearance’ features target space. Like in section (i), a pair of substructures was considered a hit if and only if the all the atoms of the query mapped all the atoms in the target.

In this approach, the low clearance reactant and the high clearance product of any transformation found were linked exclusively by way of a standard medicinal chemistry operation. This can be observed in Scheme 1.c.i and Scheme 1.c.ii.

6. Reduction to meaningful and unique transformations

The reactant and product SMILES strings in each transformation were compared, and if they were found identical then the reaction was eliminated from the reaction set.

Subsequently, any redundant pair of reactant-to-product features found within each set of transformations was eliminated from the set by using the merging component in Pipeline Pilot. The merging properties were the reactant and product SMILES strings.

7. Validation with real molecular pairs

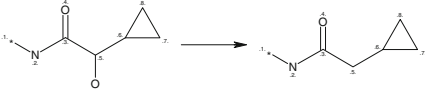
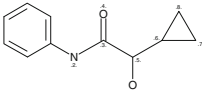
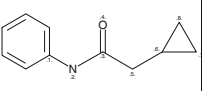
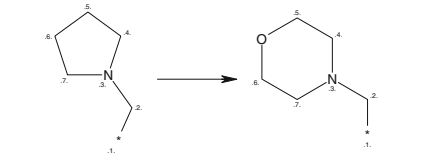
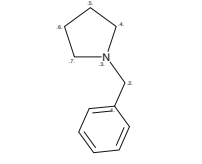
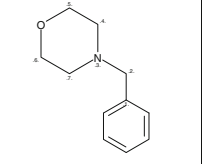
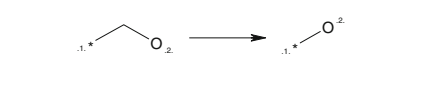
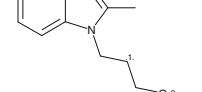
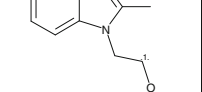

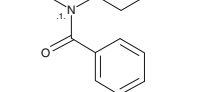
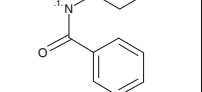
To assess the correct performance of the transformations found, the entire collection of low-to-high clearance feature pairs was run through an in-house data mining tool called ‘Buy me grease’ [12], which is designed to evaluate how specific alterations on chemical structure can affect in vitro ADME properties. The tool searched within the dataset of microsomal glucuronidation clearances for pairs of molecules that matched any given chemical transformation from the set obtained at the end of step 6. Then the ratio between the reacting and product clearances was calculated, to enable further analysis of the effect of each key transformation on the clearance of a compound.

Results and discussion

Comparison between ‘Direct SSS Transformations’, ‘Combined SSS Transformations’ and ‘Combined Exact Transformations’

At the end of step 6 in “[Materials and Methods](#)”, a total of 55,987 transformations were found, spread amongst the three different approaches as follows: 50,638 (90.5%) ‘Combined SSS Transformations’, 3,960 (7%) ‘Direct SSS Transformations’ and 1,389 (2.5%) ‘Combined Exact Transformations’. An example of each type of transformation can be seen in Table 1. It is not surprising to see such a large number of ‘Combined SSS Transformations’, since the application of a set of N traditional medicinal chemistry modifications to a feature can potentially originate N modified features, meaning that for each original feature there are potentially N-1 additional queries generated. As a consequence, there are potentially N-1 more possibilities to find a match between a query and the target

Table 1 Specific examples of transformations with matching molecular pairs

Transformation method	Reaction name	Reaction scheme	Example of matching 'reactant' molecule	Example of matching 'product' molecule	Average change (N)
Direct SSS	II - 000250				5.78 (4)
Combined SSS	I - 000735 Involving pyrrolidine_ring_break				2.65 (1)
Combined Exact	II - 001276 Involving methylene_deletion				1.31 (6)
Classical Med Chem	Demethylation				0.74 (73)

The first column contains the type of transformation method used to derive that example. The name of the specific transformation example appears in the second column. Column three shows the reaction scheme that describes the actual structural transformation. The reactants and products represented in these reaction schemes are structural motifs characteristic of compounds with either low (reactant) or high (product) glucuronidation clearance. The star symbols in the reaction schemes indicate 'any non-hydrogen atom'. Example pairs of molecules containing the reactant and the product features have been included in the fourth and fifth columns, in order to illustrate how each reaction motif can be mapped to a whole molecule. For each pair of molecules mapped to the transformation, a product:reactant clearance ratio can be obtained by dividing the clearance of the product by the clearance of the reactant. This value has been used to calculate the average change expressed in the last column. It refers to the geometric mean of the product:reactant clearance ratios obtained for all the molecular pairs that match the specific reaction. The value in brackets is the number of matching pairs to that reaction, over which the average value has been calculated

space than in a direct SSS search. At the same time, the low number of 'Combined Exact Transformations' found is also reasonable, because exact searches are much more restrictive than SSS searches. So in terms of scalability of the method, the best compromise seems to be the reciprocal feature Sub-Structural Search (SSS) method, which provides a substantial quantity of key transformations without requiring an extreme amount of processing memory.

In addition to the scalability of the process, it is very important to assess how effective the transformations are in inducing the desired change they were designed for. The outcome of step 7 in the "Materials and Methods" section was a total of 598 pairs of compounds that matched any of the 55,987 key transformations, and for which the glucuronidation clearances had been measured. Table 1 shows representative examples of how a pair of compounds matches each different type of transformation. The large difference between the number of reactions available and the number of specific examples found stems from the fact that the former are the product of an abstraction process, which makes it very unlikely to find the same amount of real examples to illustrate those changes. The proportion of

examples found for each type of transformation was in correlation with the number of transformations available for each method, thus the most successful method was the 'Combined SSS Transformations' with 434 pairs of real examples (72%), followed by 'Direct SSS Transformations' with 118 pairs (20%) and finally the 'Combined Exact Transformations' method with 46 pairs (8%).

Figure 2 shows how the ratios of product:reactant UGT clearances distribute across the three different sets of key transformations. It can be seen that the centers of gravity of the SSS methods sit well above the unity mark, whilst the pairs found for the exact search method are shifted towards the left-hand side of the plot. This indicates that the key transformations found via SSS methods can achieve higher increases in glucuronidation clearance than the exact method. The average ratios found for each type of modifications confirm this trend, with values for 'Combined Exact Transformations' (1.08) \ll 'Combined SSS Transformations' (2.01) \approx 'Direct SSS Transformations' (2.05). In addition, these ratios express the average extent of the change in clearance that can be induced by each type of key transformation: barely any change with 'Combined

Exact Transformations’ and a twofold increase with ‘Direct SSS Transformations’ or ‘Combined SSS Transformations’. A table summary of all the example transformations discussed in this section is available as Table S1 in the Supplementary Material.

Comparison of ‘Direct SSS Transformations’, ‘Combined SSS Transformations’ and ‘Combined Exact Transformations’ with pre-defined classical transformations used in medicinal chemistry

In order to assess how much of an improvement these methods are relative to traditional medicinal chemistry strategies, step 7 was expanded to the set of 440 pre-defined transformations used in steps 4 and 5 of the “Materials and Methods” section. The number of matching molecular pairs found was 984, more than double the number of examples found for the most prolific method developed in the present study, i.e. the set of ‘Combined SSS Transformations’. This is a reflection of the strategy used to derive the series contained in the dataset of in vitro glucuronidation data, for obvious reasons based on classical medicinal chemistry transformations.

The resulting distribution of product:reactant clearance ratios has also been added to Fig. 2. The centre of that distribution is a ratio of 0.95, the lowest of all the methods seen so far. This result indicates that the new methods described in the present work provide key transformations that produce a higher increase in glucuronidation clearance than the classical structural modifications. In addition, if we take into consideration that the contribution of traditional chemistry transformations to each method decreases when going from the purely pre-defined set, to the combined exact set, and then to the combined and the direct

SSS sets, a pattern seems to emerge whereby the stronger the weight of traditional medicinal chemistry transformations in the method, the least effective the structural modifications are in achieving the desired effect on the endpoint, although closer to an already well known synthetic strategy to move from reactant to product. The explanation probably stems from the fact that traditional medicinal chemistry transformations are described in a very generic way, and their combination with key features analysis limits them to the specific structural context where they work in the desired direction, i.e. that has led to successful changes in the endpoint of interest. All the example transformations discussed in this section have been included in Table S1 as Supplementary Material.

An additional analysis was performed to assess how the in silico and the medicinal chemistry transformation datasets compare when applied to an independent dataset. For this purpose, a dataset of glucuronidation clearance measures for 457 new compounds was collected. As anticipated, the number of pairs of molecules that were mapped to the transformations found via the in silico methods was rather limited (only 210 instead of the ~600 found in the original dataset), especially in comparison to the pairs found via classical medicinal chemistry transformations (571, which constitutes 73% of the total amount of pairs found in this second dataset). As explained earlier, this is a direct consequence of the current drug design culture, which has a natural tendency to use classical medicinal transformations. It is also important to remember that the in silico transformations contain moieties more specifically defined than the generic medicinal chemistry transformations, which makes the search for matching molecules much more challenging for the former type of transformations, in particular when the search is performed on an independent dataset. The results of the analysis are summarised in Table S2 and Figure S1, supplied as Supplementary Material. The average ratio of product:reactant clearance for the in silico transformations is 1.09, whilst the average ratio obtained via the classical medicinal chemistry set of transformations is 0.93. This difference is also reflected in the shape of the curves presented in the figure supplied, where a slight bias of the in silico distribution towards high values can be observed. Therefore the trend is consistent with the previous validation.

Comparison of ‘Direct SSS Transformations’, ‘Combined SSS Transformations’ and ‘Combined Exact Transformations’ with previous methods in search of effective chemical modifications

The systematic search for key structural modifications has become a topic of increasing interest in the last few years. More and more frequently, we see publications where

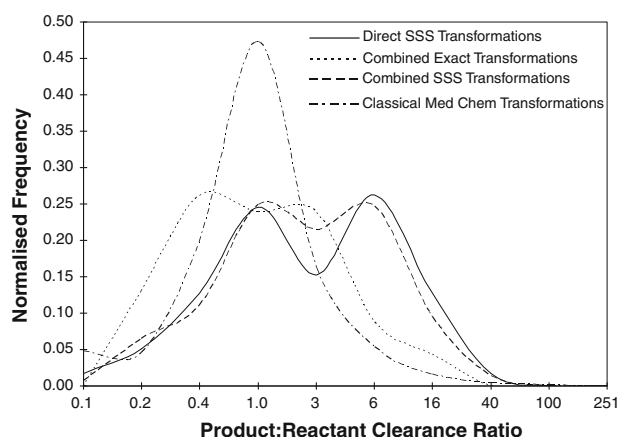


Fig. 2 Distribution of the ratios of product:reactant UGT clearances across four different sets of key transformations. It can be seen that the average values of the exact search method and the classical medicinal chemistry method are centered on the unity mark, whilst the pairs found SSS methods sit well above it

computational chemists are tackling questions of the type “given a compound on hand, what changes can I make to it to improve the activity based on the transformations within the dataset” [13], or answering “the plea from the bench chemist, ‘what do I make next?’” [14].

Although this work and those publications share the same aim of avoiding tedious manual inspection of vast molecular datasets [3, 15], a fundamental difference exists between them: primarily the use of predefined transformations in the current approach is only an option, whilst in the strategies previously published the use of predefined transformations or predefined fragmentation rules is an intrinsic requirement. Drug Guru [11] is probably the most paradigmatic example of the use of predefined transformations, as functional group and molecular framework modifications were obtained directly from medicinal chemists and encoded as pseudo reactions. Similarly, THINK and Randsmi are engines developed for the generation of new ideas for the modification of a query molecule, by applying a series of predefined transformations such as changing H to F or Me to Cl, as well as random permutations of structures using primitive operations e.g. “add a double bond” or “remove an atom” [14]. Leach also predefined modifications, involving the addition of substituents to aromatic rings and the methylation of heteroatoms [3]. And more recently Hadjuk and Lewis focused on single change predefined transformations, the former using the Daylight’s ‘findsub’ algorithm [16], and the latter resorting to classical medicinal chemistry modifications on phenyl derivatives [12]. As for fragmentation rules, the most prolific area is the search for bioisosteric replacements, which has originated a number of reviews in the literature [17–21]. The most recent bioisosteric study was carried out by Wagener, who developed a new bioisostere search method based on R-groups, linkers and cores, corresponding to fragments with one, two or three attachment points respectively [17]. Wolohan identified promising alternative core structures for lead optimization performing a Structural Unit Analysis (SUA) of fragments obtained by removing rotatable bonds to non-rotatable parts of the molecule [15], in a similar fashion as RECAP fragments are generated [22]. Other methods of predefined fragmentation rules employed to obtain key modifications include the use of Maximum Common Sub-Structure (MCSS) mappings [13, 23, 24] and extraction of side-chains [25] as described by Lewell [26]. All these methods have successfully analysed structural data in a format that parallels the intuition of medicinal chemists, making the proposed chemical modifications highly feasible. From this point of view, the approach taken here is clearly in disadvantage. However, the method here presented offers a number of advantages worth considering, which are described in detail below.

Since entering a list of pre-defined transformations or rules in the algorithm is not an imperative of the method herewith described, this approach becomes very suitable to constant automatic updates—something relatively easy to implement with for example Pipeline Pilot scheduled tasks [4]. In contrast to the studies mentioned above, the definition of ‘building blocks’ in the present analysis is rather fluid, only limited by the size of SciTegic fingerprints used, which in this case study consist of fragments containing concentric paths that range between 2 and 6 atoms long. The immediate consequence is that the analysis becomes completely naïve and objective, without any bias as to which sort of transformations or monomers have worked in the past for a specific therapeutic area, project or even research scientist. This is particularly true when the Direct SSS Transformations are derived (see point 3 in “[Materials and Methods](#)”) and, as the ‘combined’ approaches illustrate (see points 4 and 5 in “[Materials and Methods](#)”), is not at all incompatible with the incorporation of user-defined transformations—for instance, to bias the set towards chemically feasible structural modifications, or to restrict the transformations to a set that has proven to work in the optimisation of some other endpoints of interest, such as potency.

There are a number of advantages that follow from the intrinsic lack of bias of this method. Firstly, the search for pairs of key substructural features is exhaustive. As the process grows quadratically with the number of compounds, this may turn into a disadvantage, although not to the extreme of Sheridan’s process, which can vary as steeply as N^4 , where N is the number of molecules in the dataset [13]. In addition, the lack of preconceptions in the search for effective modifications pushes the SAR towards areas that have not been previously and conscientiously explored, thus opening the eyes of the scientist to new design possibilities. Such a method can offer very interesting and valuable modifications of the all-important patent space, as recognized by Wolohan [15] and Southall [23], highlighting non-classical replacements that are potentially outside of the intellectual property of a competitor or even one’s own company’s. This advantage is further enhanced by the fact that pair wise comparisons are performed between fragments with a wide range of sizes, rather than fragments of fixed sizes such as whole molecules [12, 16] or maximum common sub-structures [13, 24]. The meaning of this is that a relationship between a bad and an improved substructure can be found even though the original dataset might not contain any pair of whole molecules illustrating that relationship, because the relationship has been found through an abstraction process (see previous section about Comparison between ‘Direct SSS Transformations’, ‘Combined SSS Transformations’ and ‘Combined Exact Transformations’).

The focus in previous attempts to identify key transformations has been to organize existing data, finding the best way to breakdown SAR into manageable structural components, with the aim to detect via an ulterior analysis the structural changes that are most impactful on an endpoint that needs to be optimized. The only exception to this approach is the Drug Guru system, where the transformations are not mapped to any particular endpoint and admittedly do not offer any guarantee to achieve the desired endpoint [11]. The present approach works in the opposite direction, i.e. the first step is to detect (via in silico QSAR techniques) the structural features that are most impactful on the endpoint that needs to be optimized, and then those features are organized into pseudo reactions that by definition should lead to an improved chemical space. This makes the search of matching sub-structural units leading to a pseudo reaction a very efficient process, because only the relevant structural features are considered. It also implies that the generation of a ‘success frequency’ value for each transformation is not necessary, as the successful fragments have been automatically selected by virtue of their normalized probability and their Bayesian scoring. In other words, from the very start the focus of the method is on optimizing a clearly targeted pharmacological or ADME profile, and this guarantees a high success rate of the pseudo reactions found, as the validation with real molecular pairs indicates (see point 7 in “[Materials and Methods](#)”).

Perhaps the major limitation of this method is its inability to detect 5- to 6- and 6- to 5-membered ring changes, or atom type substitutions such as nitrogen by carbon or vice versa. The method is not focused on single-step modifications nor changes that are synthetically easy to incorporate in a series, and is not designed to detect successful monomer replacements in generic reaction schemes. So unless combined with predefined transformations as shown in point 5 of “[Materials and Methods](#)”, the modification rules resulting from this analysis will often have to be considered idea generators, rather than specific and conclusive solution generators. In addition to this, when the rules are applied there is no guarantee that the modified molecule will maintain its number of chiral centers, or that more than two heteroatoms will be positioned next to each other. This is type of limitation has been noted in previous methods [11, 14], and the solution has been suggested that the modified structures are subject to some sort of ulterior filtering or ranking method.

Conclusions

A new method to identify key chemical structure modifications based on the concept of reverse in silico QSAR has

been presented. The method has been developed by combining Naïve Bayesian statistics with SciTegic FCFP₆ molecular fingerprints. It has been exemplified with a case where changes in chemical structure that lead to a significant increase of UGT glucuronidation clearance were sought in order to improve the ADME profile of compounds for inhaled administration.

It has been shown that the new method overcomes the efficiency of traditional medicinal chemistry transformations even when the latter are inserted in the algorithm, as the respectively obtained twofold average increase versus no increase in UGT clearance values demonstrate.

This work provides a tool capable for the first time to identify chemical transformations with a focus to optimize a particular endpoint. The lack of restrictions to pre-defined transformations or fragmentation rules leads to unbiased and novel solutions for the drug design process, thus enabling a fast expansion beyond the current patent space. In summary, this method opens the door to a completely new paradigm in the drug design process.

Acknowledgments The author thanks Marcel de Groot and Willem van Hoorn for useful discussions about pair wise comparisons and Bayesian statistics. She also thanks Chad Stoner, Inaki Morao and Willem van Hoorn for reviewing the manuscript and for their helpful suggestions.

References

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46:3–26
2. Martin YCA (2005) Bioavailability score. *J Med Chem* 48:3164–3170
3. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B (2006) 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* 49:6672–6682
4. SciTegic, Inc. (a wholly-owned subsidiary of Accelrys, Inc.), 10188 Telesis Court, Suite 100, San Diego, CA 92121-4779, USA, Pipeline Pilot 7.5, 2008, version 7.5, www.scitegic.com
5. Uchaipichat V, Winner LK, Mackenzie PI, Elliot DJ, Williams JA, Miners JO (2006) Quantitative prediction of in vivo inhibitory interactions involving glucuronidated drugs from in vitro data: the effect of fluconazole on zidovudine glucuronidation. *Br J Clin Pharmacol* 61:427–439
6. Xia X, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. *J Med Chem* 47:4463–4470
7. Rogers D, Brown RD, Hahn M (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J Biomol Screen* 10:682–686
8. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46:1124–1133

9. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model* 46:462–470
10. Sun HA (2005) Naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem* 48:4031–4039
11. Stewart KD, Shiroda M, James CA (2006) Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg Med Chem* 14:7011–7022
12. Lewis M, Cucurull-Sanchez L (2009) 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* 23:97–103
13. Sheridan RP, Hunt P, Culberson JC (2006) Molecular transformations as a way of finding and exploiting consistent local QSAR. *J Chem Inf Model* 46:180–192
14. Lewis RAA (2005) General method for exploiting QSAR models in lead optimization. *J Med Chem* 48:1638–1648
15. Wolohan PRN, Akella LB, Dorfman RJ, Nell PG, Mundt SM, Clark RD (2006) Structural unit analysis identifies lead series and facilitates scaffold hopping in combinatorial chemistry. *J Chem Inf Model* 46:1188–1193
16. Hajduk PJ, Sauer DR (2008) Statistical analysis of the effects of common chemical substituents on ligand potency. *J Med Chem* 51:553–564
17. Wagener M, Lommerse JPM (2006) The quest for bioisosteric replacements. *J Chem Inf Model* 46:677–685
18. Patani GA, LaVoie EJ (1996) Bioisosterism: a rational approach in drug design. *Chem Rev* 96:3147–3176
19. Olesen PH (2001) The use of bioisosteric groups in lead optimization. *Curr Opin Drug Discov Devel* 4:471–478
20. Wermuth CG, Camille GW (2003) Molecular variations based on isosteric replacements. In *the practice of medicinal chemistry*, 2nd edn. Academic Press, London, pp 189–214
21. Lima LM, Barreiro EJ (2005) Bioisosterism: a useful strategy for molecular modification and drug design. *Curr Med Chem* 12:23–49
22. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) 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* 38:511–522
23. Southall N, Ajay T (2006) Kinase patent space visualization using chemical replacements. *J Med Chem* 49:2103–2109
24. Sheridan RP (2002) The most common chemical replacements in drug-like compounds. *J Chem Inf Comput Sci* 42:103–108
25. Haubertin DY, Bruneau PA (2007) Database of historically-observed chemical replacements. *J Chem Inf Model* 47:1294–1302
26. Lewell XQ, Jones AC, Bruce CL, Harper G, Jones MM, McLay IM, Bradshaw J (2003) Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J Med Chem* 46:3257–3274