Applying Medicinal Chemistry Transformations and Multi-parameter Optimization to Guide the Search for High Quality Leads and Candidates

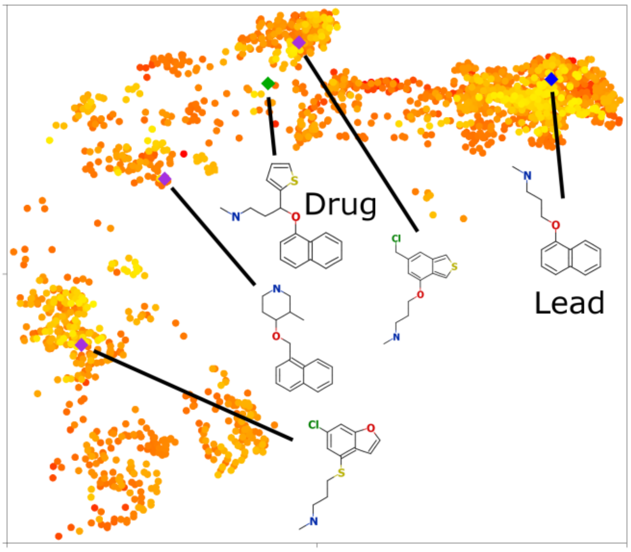
Matthew Segall[†*], Ed Champness[†], Chris Leeding[†], Ryan Lilien[‡], Ramgopal Mettu[‡], Brian Stevens[‡]

[†] Optibrium Ltd., 7226 IQ Cambridge, Beach Drive, Cambridge, UK
[‡] Cadre Research Labs, Acton, MA, USA

[*]Corresponding author. Email matt.segall@optibrium.com, Tel. +44 1223 815902

Abstract

In this article we describe a computational method that automatically generates chemically relevant compound ideas from an initial molecule, closely integrated with *in silico* models and a probabilistic scoring algorithm to highlight the compound ideas most likely to satisfy a user-defined profile of required properties. The new compound ideas are generated using medicinal chemistry 'transformation rules' taken from examples in the literature. We demonstrate that the set of 206 transformations employed is generally applicable, produces a wide range of new compounds and is representative of the types of modifications previously made to move from lead-like to drug-like compounds. Furthermore, we show that more than 94% of the compounds generated by transformation of typical drug-like molecules are acceptable to experienced medicinal chemists. Finally, we illustrate an application of our approach to the lead that ultimately led to the discovery of Duloxetine, a marketed serotonin reuptake inhibitor.



Preprint of paper to appear in J. Chem. Inf. Model. http://dx.doi.org/10.1021/ci2003208

Introduction

In silico predictive models of key properties are routinely used in the selection and design of potential drug molecules[1]. These results may be combined to prioritize compound ideas for synthesis, simultaneously optimizing multiple parameters to identify compounds with an appropriate balance of properties for the therapeutic goal of a drug discovery project[2,3]. Furthermore, the structure-activity relationships that these models capture can guide the redesign of compounds to improve their properties and overcome liabilities[4].

Predictive methods can score and rank compounds to guide the search for high quality compounds among a large number of possibilities; therefore, getting the maximum value depends on having a rich set of potential compounds to search. However, during optimization it is rare for a large library of relevant, predefined structures to be available and it is common to rely on a medicinal chemist to define possible compounds of interest, either by drawing individual structures or enumerating virtual libraries based on a common structural motif. This is a time consuming process and limited by the experience of an individual chemist.

Methods for automatically applying medicinal chemistry 'transformation rules' to generate new compound structures have been previously described[5,6]. These typically accept an initial 'parent' structure as input and generate 'child' structures by applying transformations based on collective medicinal chemistry experience. Examples of transformation rules range from simple substitutions or bioisostere replacements to more dramatic modifications of the molecular framework such as ring opening or closing. A computer can store and apply many more rules than a single chemist and can 'learn' from historical examples of transformations between molecules[7]. Applying a set of transformations iteratively to produce multiple 'generations' of compound ideas can result in a large number of molecules – too many to be examined visually by a chemist to select the most interesting for further consideration.

In this paper, we describe the combination of an algorithm to generate compound ideas, by applying transformations to an initial molecule, with predictive models and a multi-parameter scoring algorithm to quickly focus attention on those ideas most likely to satisfy the required property profile. The goal is a tool to support experts and stimulate the process of innovation – achieving a creative combination of a computer's ability to cover a wide breadth of possibilities with the experience and detailed knowledge of a chemist. In particular, the discovery process should be directed by an expert and provide a prioritized list of possibilities for further consideration, not an automatically designed final compound.

To be successful, such a method must satisfy a number of requirements:

- It must generate a wide diversity of chemistry, as the objective is to explore many ideas in the search for an optimal solution.
- The compound structures generated must be relevant. In particular, the number of 'nonsensical', e.g. chemically unstable or infeasible, compounds must be kept to a minimum. Also, the chemist must be able to control the generation process, for example by specifying a region that must not be modified or restricting the transformations that will be applied.
- The transformations that are applied should include a broadly representative set of those applied successfully in the past to optimize successful drugs.
- The method used to prioritize the resulting compound ideas should reliably identify high quality compounds within those given the highest rank in the generated set.

The methods used to create and apply a set of transformations and prioritize the compounds generated thereby are described in the Methods section. In the Results section, we describe the validation of this method to ensure that the transformations cover a broad range of 'drug like' chemistry and that the resulting structures are relevant and not unstable or infeasible. We will describe the application of our method to efficiently identify compounds similar to known drugs, starting from the lead compounds from which the drugs were derived. Furthermore, we illustrate the application of the compound idea generation method combined with predictive models and a multi-parameter optimization algorithm to the lead of a known drug, Duloxetine. Although retrospective, this application will demonstrate the ability to efficiently target high quality compounds. Finally, we will discuss possible applications of these methods and draw some conclusions.

Methods

Transformations

Two hundred and six transformations were generated by study of medicinal chemistry literature[8-24] and observation of the optimization steps between known drugs and the lead molecules from which they were derived.

The transformations were divided into seven broad groups: Functional Group Addition, Linker Modification, Remove Atom, Ring Addition, Ring Modification, Ring Removal, Terminal Group Exchange. The distribution of transformations between the groups is shown in Table 1 and examples of each are shown in Table 2.

Table 1 Distribution of transformations between groups.

Group	Number of transformations
Functional Group Addition	20
Linker Modification	54
Remove Atom	5
Ring Addition	13
Ring Modification	26
Ring Removal	4
Terminal Group Exchange	84
Total	**206**

The transformations do not necessarily correspond to specific chemical reactions or synthetic routes; rather they are intended to describe changes to molecules that a medicinal chemist might consider in the course of an optimization project. A single transformation might require multiple synthetic steps or the synthesis of new building blocks. However, the transformations are typically not major rearrangements – they are relatively feasible moves in chemical space.

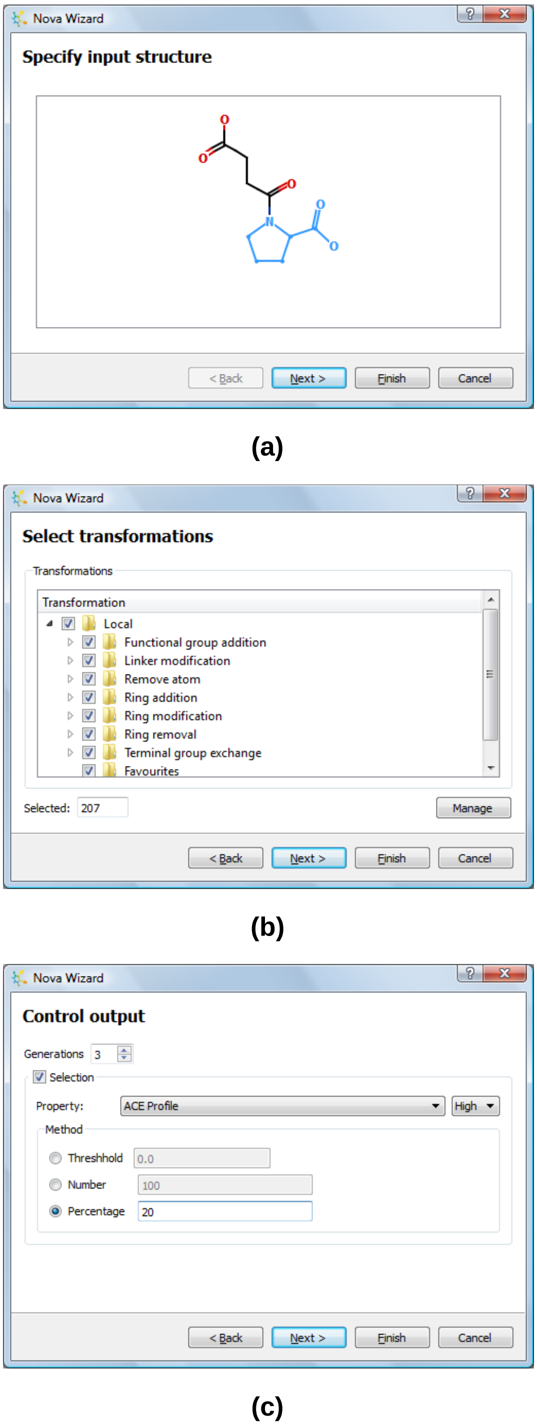
Representation of Transformations

The compound transformations were encoded as SMIRKS, a reaction transform language designed by Daylight Chemical Information Systems which uses SMILES and SMARTS notations to specify a generic reaction or transformation[25]. SMIRKS representations of example transformations are provided in Table 2.

Generation of Compound Structures

The Cactvs cheminformatics library[26] was used within the StarDrop software platform[27] to apply the transformations to a parent compound structure encoded as a SMILES string. The Cactvs implementation also allows a fragment of the parent to be specified as a SMARTS pattern, such that this fragment will not be modified during the generation process and any transformations that would modify this region will be ignored.

The user can specify the parent structure and control the generation process through a graphical user interface. The typical workflow is illustrated in Figure 1: The user can specify a region of the compound that must not be modified; the transformations to be applied can be selected; the number of generations of transformations to be applied can be specified; and finally, because this process generates a number of compounds that grows exponentially with the number of generations, the user can control this growth by specifying a criterion to select a subset of the compounds in each generation. The criterion may be defined in terms of any predicted property or a score that represents the overall quality with respect to a profile of properties (see "Scoring" below) and can be specified as a threshold value for the property, e.g. only accept compounds with logS > 1, or the number or proportion of compounds to select from a list ranked by the property, e.g. only progress the 100 compounds with the highest predicted potency in a generation or the highest scoring 10% of a generation.



Nova Wizard

Specify input structure

(a)

Nova Wizard

Select transformations

Transformations

Transformation

- ☑ Local
  - ☑ Functional group addition
  - ☑ Linker modification
  - ☑ Remove atom
  - ☑ Ring addition
  - ☑ Ring modification
  - ☑ Ring removal
  - ☑ Terminal group exchange
- ☑ Favourites

Selected: 207

(b)

Nova Wizard

Control output

Generations 3

☑ Selection

Property: ACE Profile High

Method

○ Threshhold 0.0

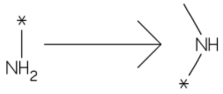
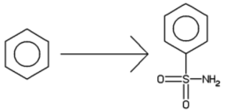
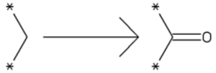
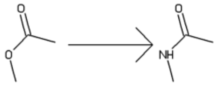
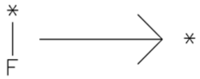
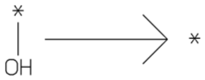
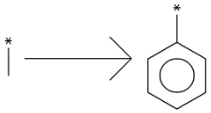
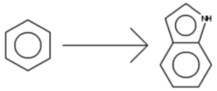
○ Number 100

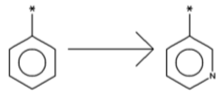
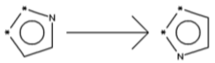
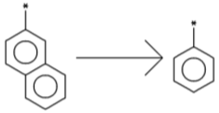
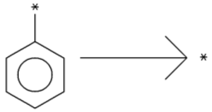
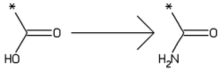
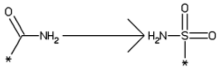
● Percentage 20

(c)

Figure 1 Illustration of a workflow to initiate the generation of new compound structures. (a) Specify the input structure. A region of the molecule can be chosen to be 'frozen' (shown in light blue), in which case no modifications will be made to this region. (b) The transformations to apply can be selected, either individually or as groups. The groups can be managed to create groups tailored to specific objectives or to add new transformations. (c) The number of generations can be specified and a criterion for selection can be defined to limit the growth of the number of compounds generated. The selection can be defined as a minimum threshold for a property or score or a maximum number or percentage of each generation that will be used as the basis for subsequent generations.

Table 2 Example Transformation Rules.

Group	Transformation Name	Illustration	SMIRKS
Functional Group Addition	Methyl addition to amine		[N:1][H]>>[N:1]C
	Sulfonamide addition to benzene		[c:1]1[c:2][c:3][c:4][c:5][c:6]1[H]>>[c:1]1[c:2][c:3][c:4][c:5][c:6]1S(N)(=O)=O
Linker Modification	Secondary carbon to carbonyl		[*;!#1:1][CH2][*;!#1:2]>>[*;!#1:1]C(=O)[*;!#1:2]
	Ester to amide linker		[#6:1]O[C;!R:3](=O)[#6:2]>>[#6:1]N[C;!R:3](=O)[#6:2]
Remove Atom	Remove halogen		[C,c:1][F,Cl,Br,I]>>[C,c:1]
	Remove hydroxyl		[C,c:1][OH]>>[C,c:1]
Ring Addition	Methyl to phenyl		[*;!#1:1][CH3]>>[*!#1:1]c1ccccc1
	Benzene to indole		[c:1]([H])1[c:2]([H])[a:3][a:4][a:5][a:6]1>>[C:1]12[a:6]=[a:5][a:4]=[a:3][C:2]=1[nH]C=C2

Ring Modification	Phenyl to 3-pyridine		[*;!#1:1][c:2]1[c:3][c:4][c:5][cH][c:6]1>>[*;!#1:1][c:2]1[c:3][c:4][c:5][n][c:6]1
	NC-switch		[*:1]1:[c]([*:2]):[c:10]([*:3]):[n]([*:4]):[*:5]1>>[*:1]1:[n]([*:2]):[c:10]([*:3]):[c]([*:4
Ring Removal	Napthalene to benzene		[*;!#1:7][c:1]1[cH]c2c([cH][c:6]1)[c:5][c:4][c:3][c:2]2>>[*;!#1:7][c:1]1[c:2][c:3][c:4][c:5][c:6]1
	Remove phenyl		[*;!#1:1]c1[cH][cH][cH][cH][cH]1>>[*;!#1:1]
Terminal Group Exchange	Carboxyl to amide		[*;!#1:1][C:2](=O)[OH]>>[*;!#1:1][C:2](=O)N
	Amide to sulfonamide		C(=O)([NH2])[*;!#1:1]>>S(=O)(=O)([NH2])[*;!#1:1]

Predictive Models

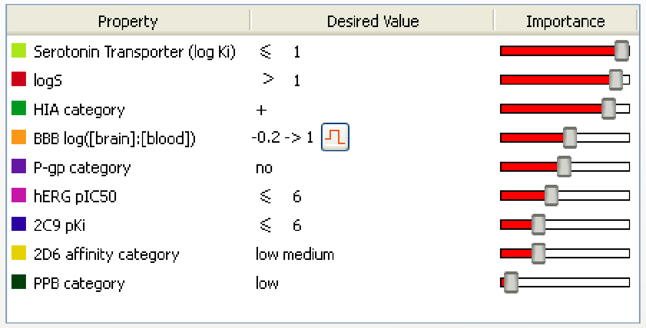
Any *in silico* model may be used to predict the properties of the compounds generated. However, due to the large number of compounds that may be generated, the models should be capable of generating predictions quickly in order to prevent the process from becoming intractable.

In the example presented in this paper, quantitative structure-activity relationship QSAR models implemented in the StarDrop software platform were used[27] to predict the following ADME and physicochemical properties: octanol/water partition coefficient (logP), aqueous solubility (logS), human intestinal absorption (HIA), blood-brain barrier penetration (logBB), inhibition of the potassium ion channel encoded by the human ether-a-go-go related gene (hERG pIC_{50}), human plasma protein binding (PPB), inhibition of cytochrome P450 isoforms CYP2D6 and CYP2C9 (pKi) and active transport by P-glycoprotein (P-gp).

In order to identify high quality compounds it is also necessary to predict activity against the pharmacological target for the intended drug. In the example application described herein, a QSAR models of the inhibitory constant against the serotonin transporter (expressed as the logarithm of the K_i in nM) was generated. The data set used to build this model was derived from the publicly accessible ChEMBL database provided by the European Bioinformatics Institute[28]. A training set of 1454 compounds was used to build multiple models using a range of statistical fitting methods using the StarDrop Auto-Modeller[29] and the model with the highest coefficient of determination (R^2) on an independent validation set of 311 compounds was selected. The resulting model used a Gaussian Processes (GP) method[30] and 62 descriptors, including logP, McGowan's volume[31], topological polar surface area[32] and two-dimensional structural descriptors defined as SMARTS patterns. The final model has an R^2 of 0.88 and root mean square error (RMSE) of 0.62 on the training set, an R^2 of 0.72 and RMSE of 0.85 on the validation set and an R^2 of 0.81 and a root mean square error of 0.76 on a further, external test set of 311 compounds. The model also estimates the confidence in each prediction, based on the GP method which relates the uncertainty in the prediction for each compound to its proximity to the compounds in the training set. This confidence is explicitly taken into account in the scoring method discussed below, so that highly uncertain predictions are not given undue weight in the selection of compounds.

Scoring

The methods underlying the probabilistic scoring algorithm employed herein are discussed in more detail in references[3,4] but here will give a brief overview. A probabilistic score is one which indicates the probability of success of a molecule against a 'scoring profile' that defines criteria for the properties that are required in an ideal compound. It is also important to specify the relative importance of the criteria as, in practice, it is often necessary to make a trade-off between properties if an ideal molecule cannot be identified. Furthermore, more subtle trade-offs can be defined than simple pass/fail criteria, as a scoring profile could contain more complex functions for each property representing a range of acceptability over the property value range. An example of such a scoring profile is shown in Figure 2.



Property	Desired Value	Importance
🟩 Serotonin Transporter (log Ki)	⩽ 1	
🟥 logS	> 1	
🟩 HIA category	+	
🟧 BBB log([brain]:[blood])	-0.2 -> 1	
🟪 P-gp category	no	
🟪 hERG pIC50	⩽ 6	
🟦 2C9 pKi	⩽ 6	
🟨 2D6 affinity category	low medium	
🟩 PPB category	low	

Figure 2 The scoring profile used to prioritise compounds generated from the Duloxetine lead, showing the properties of interest, the desired value ranges and the importance of each criterion. For example, the most important property was inhibition of the serotonin transporter, for which a predicted Ki of less than 10 nM (log K_i <1) was required. This was followed by an aqueous solubility of greater than 10 μM (logS > 1) and positive prediction for human intestinal absorption.

When combining property data on multiple properties, it is also important to consider the uncertainty in each data point, as this could lead to the overall uncertainty in the scores being high, reducing our ability to confidently distinguish high and low quality molecules. The result of this process is a score for each molecule, representing the likelihood of a molecule meeting the scoring criteria and an uncertainty in the overall score, derived from the uncertainties in each of the individual property values. These uncertainties can be used to establish whether the available data allow one molecule to be confidently chosen over another.

Similarity

Compound similarity was measured using the Tanimoto index calculated between topological path-based fingerprints, with a maximum path length of 7 and a fingerprint size of 2048 bits. This was performed using the RDKit toolkit[33].

Drug Data Set

The set of 3,211 drug molecules used in the validation of the transformations (the 'drug set') was derived as follows: Version 2.5 of the DrugBank Small Molecule database[34] was obtained on August 23, 2010. This initial set containing 4854 molecules was reduced by removing molecules containing atoms other than C, H, N, O, P, S, Cl, or F, molecules with molecular weight less than 200 Da and 140 molecules which contained poorly specified SMILES (127 aromaticity errors and 13 valence errors), resulting in 3214 compounds. Finally, three additional molecules (insulin, inulin and DB05413) were removed, as these are very large, not representative of the compounds to which we expect this method to be applied and likely to skew the validation statistics due to their size. 40 compounds were slightly edited to remove small cofactors or counter-ions or to select only one isomer where multiple isomers were specified.

Results

Transform Set Validation

Coverage

In order to ensure that the set of transformations employed covers a wide range of 'drug-like' chemistry, enabling the exploration of a diverse range of potential modifications, each transformation should apply to a wide range of molecules; a transformation that uniquely applies to a single molecule is not of interest. Furthermore, when the full set of transformations is applied to a 'typical' drug-like parent molecule, a large number of child molecules should be generated.

To test these requirements, the 206 transformations were applied to a set of 3211 drug molecules – the 'drug set' described in the Methods section. This resulted in 584,124 child compounds; thus, on average, 182 child compounds were generated from each parent. Furthermore, on average, each transformation applied at least once to 31% of the molecules in the drug set.

These statistics indicate that the set of transformations have broad applicability to drug-like compounds and will generate a wide range of child compounds.

Quality

As discussed above, the transformation rules should be sufficiently general. However, there is a trade-off in that a more general transform is more likely to apply in an occasionally inappropriate chemical context. This can generate undesirable or infeasible compound structures. The desirability of compound structures is, to some extent, subjective. Therefore, the quality of the compound structures generated was assessed by asking two independent medicinal chemists to examine a set of 1,500 compounds generated using the 206 transformations.

The quality assessment set was generated as follows: 400 compounds were randomly selected from the drug set described in the Methods section. All of the 206 transformations were applied to the 400 selected molecules to generate a set of child compounds. From the full set of child compounds, 1500 were selected at random for assessment by the medicinal chemists.

The medicinal chemists were asked to assess each child compound to determine whether it was undesirable. They were not asked to determine if they could identify a synthetic route to the product – an ideal compound that was synthetically challenging may be worth the effort of devising a difficult synthetic route or may spark further ideas that are more accessible.

From the same set of 1500 child compounds, one chemist flagged 7% of the structures as undesirable while the other flagged 4.1%. This demonstrates that desirability is, to some extent, subjective. However, an average acceptance rate of 94% was considered to be more than sufficient. It would be possible to filter out some of the undesirable structures before they are output. However, it was decided to retain this small proportion of poor compound structures, though they may be a minor distraction, as they may stimulate ideas for similar compounds that are chemically feasible.

Hit-like to Drug-like Transformation Series

The transformations in the set should be representative of those used in practice to optimize leads into drug molecules. To assess this, a data set containing 60 marketed drugs and the initial leads from which they were derived, published by Perola[35], was used (we will refer to these lead/drug pairs as the 'Perola set').

For each lead/drug pair in the Perola set, the lead was used as the initial parent and the 206 transformations were applied iteratively to explore the 'universe' of compounds that are accessible from the lead. The goal of this was to identify the closest compound structure in this universe to the corresponding drug. This is challenging, as many of the derivations of drugs in the Perola set from their corresponding leads include the exchange or incorporation of large or relatively uncommon fragments. A result of the coverage requirements described above is that most of the transforms involve smaller fragments. Therefore, many iterative applications of the transformations may be required, creating many generations of child compounds, to move from a lead to a compound similar to the corresponding drug and, even then, it may not be possible to find an exact match to the drug.

As the number of compounds generated increases exponentially with the number of generations, it is impractical to exhaustively enumerate all offspring compound structures. For example, if 182 compounds are generated on average from a single parent, the third generation will contain more than 6 million compounds. Therefore, a 'beam' search was implemented, whereby the 100 compounds with the greatest similarity to the target drug were retained after each iteration and a total of five iterations were applied. The closest match, as measured by Tanimoto similarity applied to topological finerprints (see Methods Section), to the corresponding drug was identified from the resulting child compounds. The disadvantage of this approach is that it does not guarantee to find the closest match that could be achieved, as it may be necessary to initially move away from the drug in order to ultimately generate the most similar compound. Furthermore, it may be possible to find a closer child compound if more than five iterations were applied.

On average, the similarity of the drug with closest match in the child compounds generated from the corresponding lead was 0.86 compared with an average similarity between the drugs and leads of 0.64. Out of the 60 Perola lead/drug pairs, nine exact matches were achieved within the compounds generated from the initial lead. The structures of the initial leads, corresponding drugs and closest identified child compounds are provided in the Supporting Information. It should be noted that this is not an external validation of the transformation set, as a few known drugs (including some from the Perola set) influenced some of the commonly applied transformations. However, this test provides confidence that the transformations chosen in the set of 206 are not only generally applicable, but can move from lead-like to drug-like compounds across a wide range of small molecule drug classes.

Example Application

To illustrate the application of the transformation set to guide the search for optimized compounds based on an initial lead, we used the lead molecule that ultimately gave rise to the drug Duloxetine as the parent molecule.

QSAR models of absorption, distribution, metabolism and elimination (ADME) properties and the inhibitory constant K_i for the serotonin transporter, described in detail in the Methods section, were used to predict the properties of the compound ideas generated. These ideas were prioritized against the multi-parameter profile of property criteria shown in Figure 2, which combines potency against the primary target with suitable ADME properties for an orally dosed compound against a CNS target. To achieve this a score between zero and one was calculated for each compound, using a probabilistic scoring method described in the Methods Section.

The application of one generation of transformations produced 172 child compounds, which suggested that exhaustive enumeration of more than two generations would be intractable. Therefore, three generations were applied, but only the top-scoring 10% of the compounds in each generation were used as the basis for subsequent generations.

The resulting data set contained 2,208 compounds and the scores for these compounds are plotted in Figure 3. From this, a number of observations may be made: First, as the results from multiple uncertain predictions are combined to calculate the scores, the uncertainties in the scores are high, as shown by the error bars in Figure 3. Therefore, it is difficult to discriminate between compounds with confidence, particularly in the later generations. However, despite this, the information provided by the score is sufficient to guide a consistent improvement and the compounds in each generation typically show an increase in score over the previous generation; the score for the initial lead is 0.09 and the averages for the compounds in subsequent generations are 0.32, 0.44 and 0.53 respectively (note that only the top 10% of the compounds in each generation are included). Furthermore, the score of the top compounds (0.7 ± 0.3) suggest ~95% confidence that they are better than the initial lead (0.1 ± 0.2), assessed against the criteria defined in the scoring profile. Finally, it is notable that Duloxetine itself is present in the final generation, with a score (0.5 ± 0.3) that is higher than the initial lead with ~90% confidence and not significantly below that of the highest scoring compounds.

The structures and scores of the initial lead and Duloxetine are shown in Figure 4 along with the three highest ranking molecules generated. Although none of the top-three compounds could be identified in a search of PubChem[36], the second-ranked compound bears a strong similarity (Tanimoto similarity >0.9) to Litoxetine, shown in Figure 4, which was progressed to clinical trials and is active against the serotonin transporter with an IC_{50} of 6 nM[37].

It is notable that the third-ranked compound in Figure 4 is likely to be an alkylating agent. This illustrates that, while we have tried to minimize the number of 'nonsensical' compounds generated by the transformations, some compounds may be generated with undesirable functionalities and we will discuss this further in the conclusions below.

A 'chemical space' visualization, illustrating the diversity of the compounds in the generated data set, is shown in Figure 5. This plot was generated by generating the full similarity space for the set of 2208 compounds, using 2D path-based fingerprints and a Tanimoto similarity index and plotting the first two principal components. From this it is notable that a wide range of different chemical motifs have been explored and that there are multiple 'hot spots' containing high-scoring compounds; the best scoring compounds are not concentrated in one region, indicating that the algorithm has identified a number of different chemical strategies worthy of further consideration. The top three ranked molecules are structurally diverse, within the range of diversity explored around the initial lead, and are distinct from both the initial lead and Duloxetine itself.

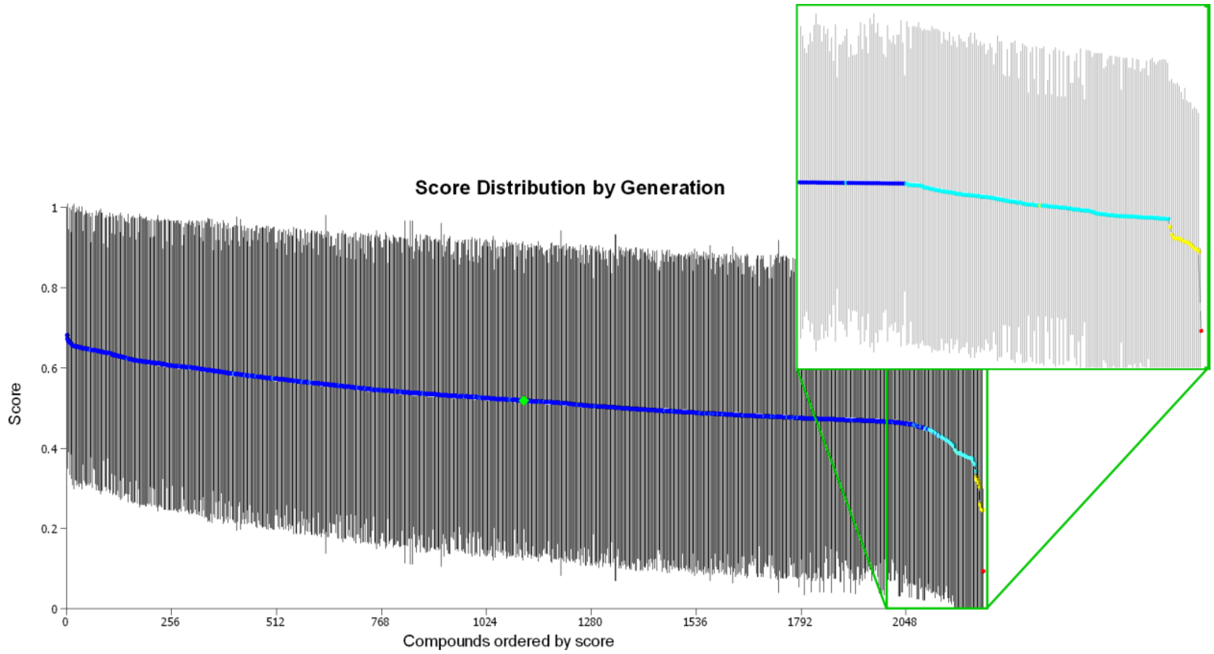


Figure 3 This graph shows the 2208 compounds generated by three generations of transformations starting with the lead compound for the project that yielded the drug Duloxetine. The compounds generated are ordered along the x-axis accoring to their score from highest to lowest and the score for each compound, as calculated using the probabilistic scoring algorithm, is plotted on the y-axis. Error bars show the uncertainty of the overall score for each compound due to the uncertainties in the underlying predictions. The compounds are coloured by generation: Red is the parent, yellow generation 1, light blue generation 2 and dark blue generation 3. The drug Duloxetine was present in generation 3 and is shown by the green diamond.

In this example, the increase in score is driven primarily by the improvements in predicted target affinity between generations because the predicted ADME properties of the lead compound were good to begin with. However, the use of probabilistic scoring to select compounds with a good balance of properties was valuable as it eliminated compounds in early generations that were predicted to have high target affinity but were unlikely to have a good balance of ADME properties for the overall objective. Figure 6 shows the distribution of the scores for compounds in the first two generation with predicted K_i less than 10 nM, indicating that a significant number of compounds that were predicted to be active were rejected due to the predictions of poor values of other properties including solubility (184 compounds from generation 2 were used as the progenitors of generation 3).

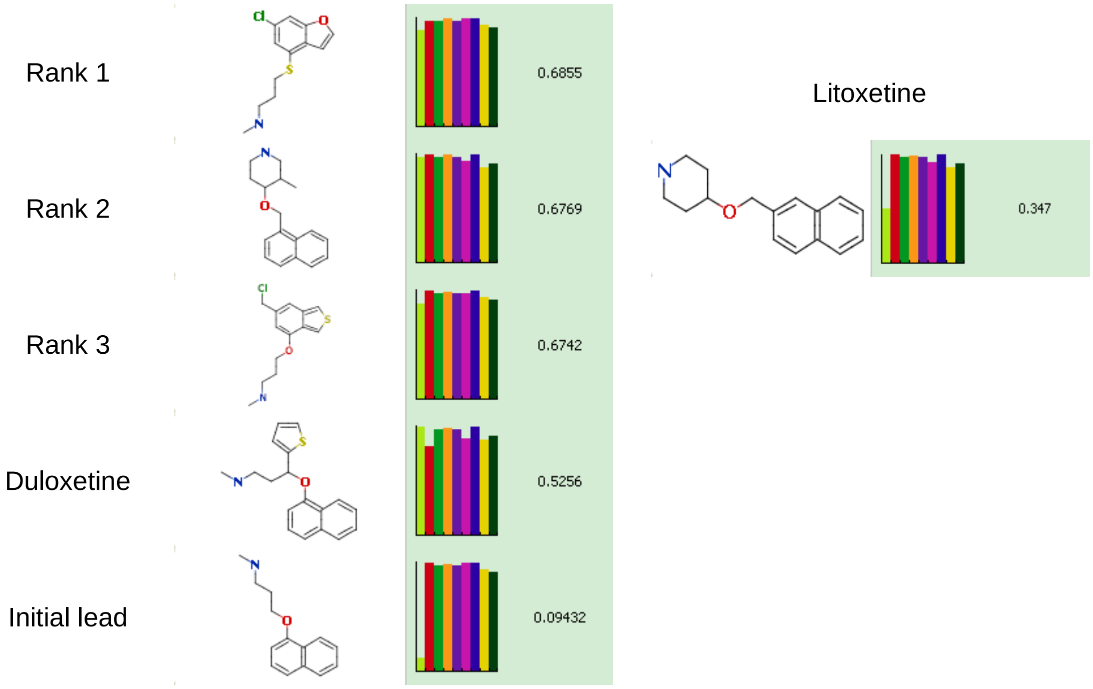


Figure 4 On the left, the initial lead that ultimately gave rise to Duloxetine, the top three compounds generated from this lead and Duloxetine, which was also generated by the algorithm are shown. The score for each compound is show to the right along with a histogram indicating the contribution of each property to the overall score (the color of each bar corresponds to the property key shown in Figure 2). For comparison with the second-ranked compound, the structure and calculated score for Litoxetine a clinical candidate serotonin reuptake inhibitor is shown on the right. Although this structure was not generated automatically in this example, it bears a strong similariy (Tanimoto similarity >0.9) with the second-ranked compound, which has a higher predicted affinity and hence a higher score.

Discussion and Conclusions

In this paper we have described an algorithm for automatically generating new compound ideas from an initial molecule using a set of medicinal chemistry transformations derived from the literature. We have shown that these transformations are generally applicable and generate structures that are relevant and acceptable to medicinal chemists. Furthermore, we have demonstrated the use of this chemical transformation algorithm coupled with predictive models and a multi-parameter optimisation method, integrated in an intuitive visual environment, to stimulate the exploration of a wide range of strategies to identify compounds with a good balance of properties and hence a high chance of downstream success.

While we use a systematic search method as the basis for making chemical modifications, other approaches based on evolutionary algorithms have also been applied[38,39] . Motivated by the theory of evolution, EA-based methods 'mutate' the structure of a compound by making small modifications to a compound structure, for example adding or removing a single atom, changing the bond order or changing a carbon atom into a heteroatom. The equivalent of genetic 'crossover' can also be implemented by combining substructures from two different compounds. This evolutionary process is guided by a 'fitness function' that may be defined in terms of simple descriptors, predicted properties or even by user selection[40]. The application of a transformation rule in our approach is analogous to a mutation; however the structural changes corresponding to medicinal chemistry transformations are typically larger than an EA mutation. Similarly, the role of a fitness function to guide the optimization in an EA is fulfilled by the probabilistic score in our approach. One advantage of a medicinal chemistry transformation-based approach is that the structures generated tend to be more relevant, due to the fact that the transformations are based on historical precedents; however, the diversity of chemistry that can be explored may be more limited than an EA approach, as it is restricted by the library of transformations applied.

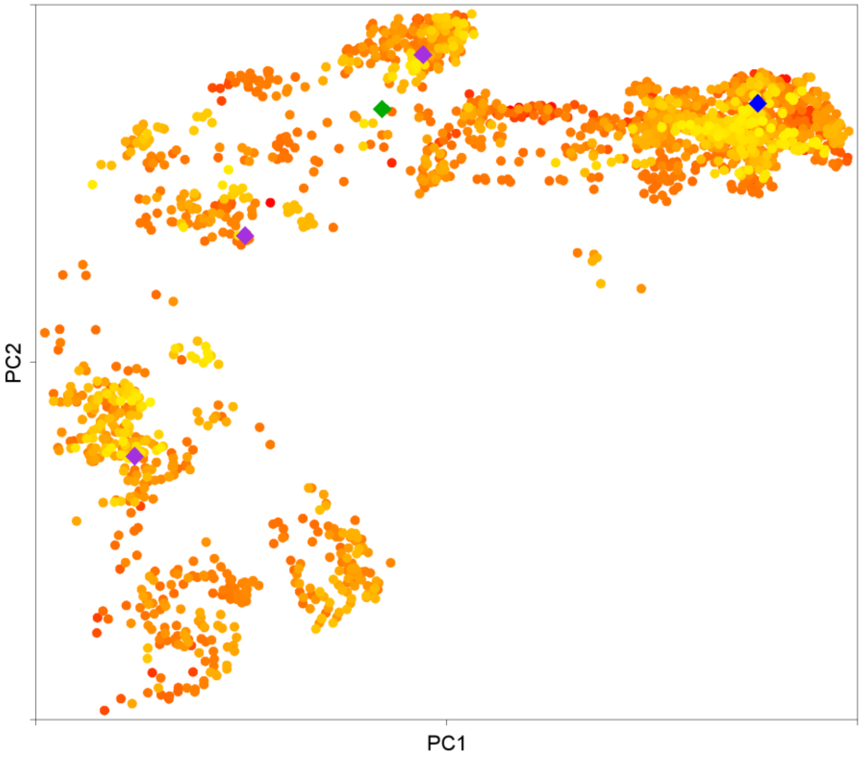


Figure 5 The chemical space of compounds generated from the initial lead that gave rise to Duloxetine. The points corresponding to compounds are coloured by score, from the lowest (0.29) in red to the highest (0.69) in yellow. The initial lead is shown as a dark blue diamond, Duloxetine as a green diamond. The top-three scoring compounds are shown as purple diamonds. In this plot, each point represents a compound and the distance between two points indicates their structural similarity; close points are structurally similar while distant points are structurally diverse. The method by which this plot was generated is described in the text.

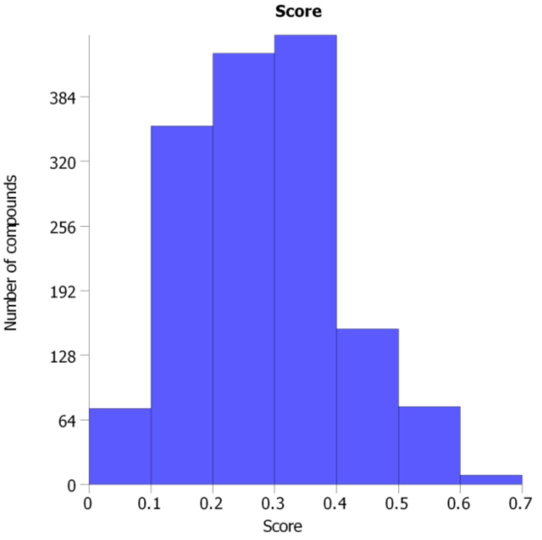


Figure 6 Score distribution for the compounds in generations 1 and 2 from the Duloxetine lead compound with a predicted K_i of less than 10 nM. From this we can see that there are a significant number of compounds with poor scores, despite having high target affinity, indicating that they are likely to have poor values for other relevant properties.

The most significant limitation of the method we have described is that, while we have shown that the large majority of the chemical structures generated are relevant and not infeasible, there is no guarantee that they can be easily synthesized from available reagents. Computational methods have been proposed for estimating synthetic tractability[41] and including such an estimate as an additional parameter in the multi-parameter optimization profile would be one approach to address this.

In the application to the Duloxetine lead, we noted an example of a compound generated with an undesirable, alkylating functionality. This is due to the fact that a transformation set should have a balance between generality and the relevance of the compounds generated; attempting to restrict the transformations to eliminate all undesirable functionalities would severely limit the diversity of chemistry that could be explored. Furthermore, a compound with an undesirable functionality may provide the seed for a valuable idea through a simple modification. However, a set of substructural alerts (e.g. [42,43]) could be applied to flag compounds that contain undesirable functionalities, either as a post-hoc filter or as a criterion in the scoring profile to de-prioritize the selection of compounds are flagged during the generation process.

Herein, we employed a two-dimensional QSAR model for prediction of potency against the therapeutic target. There are many other approaches for prediction of potency that take into account 3-dimensional information, such as pharmacophore, docking or shape-based methods. It should be noted that any method may be used to predict the properties used to calculate the probabilistic score used to guide the selection of compounds between generations and three-dimensional methods would provide a good approach to eliminate compounds that do not fit the active site of the target.

Another potential extension would be to explicitly consider diversity in the selection of compounds between generations. In the example application described herein, only the top scoring compounds were selected as the basis for subsequent generations. There is a risk that such a search strategy could quickly focus on a set of very similar compounds, although as we demonstrated this did not occur in this case. To mitigate this risk, a selection could be made based on a balance of score and structural diversity[44,3] which would select some lower-scoring compounds where these would add significantly to the diversity of the compounds selected and would prevent the exploration from becoming trapped in a local maximum. The degree of bias between score and diversity could be a user-controlled parameter.

There are a wide range of potential applications of this technology. These include: aiding the rigorous exploration of chemistry around early hits, to identify those hits most likely to yield high quality lead series; helping to find strategies to overcome problems with compound properties in lead optimisation; and identifying patent busting opportunities by expanding the chemistry around existing development candidates or drugs to search for compounds with improved properties.

Finally, while we have focused on the creation and validation of an initial set of transformations, it is possible to extend this set with new transformations based on the experience of medicinal chemists or designed around specific chemistry available within an organization. Furthermore, it may be beneficial to organise transformations into groups, perhaps tailored to specific objectives such as improving metabolic stability or reducing plasma protein binding. Thus, this approach could be used as a tool to capture and share knowledge between medicinal chemists or even as an educational resource for less experienced scientists.

Acknowledgements

The authors would like to thank Miss Delphine Lariviere for generating the model of serotonin transporter inhibition used in the example application. We would also like to thank Dr Nawaz Khan and Dr Colin Greengrass for each visually inspecting 1,500 molecules to independently assess their acceptability.

Supporting Information Available

Table of initial leads, marketed drugs and closest child compounds generated in 5 generations for Perola set. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

1. Van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003,** *2,* 192-204.
2. Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Towards a new age of virtual ADME/TOX and multidimensional drug discovery. *J. Comput.-Aided Mol. Des.* **2002,** *16,* 381-401.
3. Segall, M.; Beresford, A.; Gola, J.; Hawksley, D.; Tarbit, M. Focus on Success: Using in silico optimisation to achieve an optimal balance of properties. *Expert Opin. Drug Metab. Toxicol.* **2006,** *2*.
4. Segall, M.; Champness, E.; Obrezanova, O.; C, L. Beyond Profiling: Using ADMET models to guide decisions. *Chem. & Biodiv.* **2009,** *6,* 2144 - 2151.
5. Stewart, K.; Shiroda, M.; James, C. Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **2006,** *14,* 7011-7022.
6. Ekins, S.; Honeycutt, J.; Metz, J. Evolving molecules using multi-objective optimization: applying to ADME/Tox. *Drug Discov. Today* **2010,** *15,* 451-460.
7. Raymond, J.; Watson, I.; Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.* **2009,** *49,* 1952-1962.
8. Burger, A. *Medicinal Chemistry,* 3rd ed.; John Wiley & Sons Inc: San Francisco, 1970.
9. Bonnet, P.; Robins, R. Modulation of leukocyte genetic expression by novel purine nucleoside analogues. A new approach to antitumor and antiviral agents. *J. Med. Chem.* **1993,** *36,* 635-653.
10. Binder, D.; Hromatka, O.; Geissler, F.; Schmied, K.; Noe, C.; Burri, K.; Pfister, R.; Strub, K.; Zeller, P. Analogues and derivatives of tenoxicam. 1. Synthesis and antiinflammatory activity of analogues with different residues on the ring nitrogen and the amide nitrogen. *J. Med. Chem.* **1987,** *30,* 678-682.
11. Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K.; Reinemer, P.; Perzborn, E. Discovery of the novel antithrombotic agent 5-chloro-N-(((5S)-2-oxo-3- [4-(3-oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl)methyl)thiophene- 2-carboxamide (BAY 59-7939): an oral, direct factor Xa inhibitor. *J. Med. Chem.* **2005,** *48,* 5900-5908.
12. Patani, G.; LaVoie, E. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996,** *96,* 3147-3176.
13. Black, J.; Duncan, W.; Shanks, R. Comparison of some properties of pronethalol and propranolol. *Br. J. Pharmacol. Chemother.* **1965,** *25,* 577-591.
14. Walsh, D.; Franzyshen, S.; Yanni, J. Synthesis and antiallergy activity of 4-(diarylhydroxymethyl)-1-[3-(aryloxy)propyl]piperidines and structurally related compounds. *J. Med. Chem.* **1989,** *32,* 105-118.
15. Fournié-Zaluski, M.; Coric, P.; Turcaud, S.; Rousselet, N.; Gonzalez, W.; Barbe, B.; Pham, I.; Jullian, N.; Michel, J.; Roques, B. New dual inhibitors of neutral endopeptidase and angiotensin-converting enzyme: rational design, bioavailability, and pharmacological responses in experimental hypertension. *J. Med. Chem.* **1994,** *37,* 1070-1083.
16. Larsen, A.; Lish, P. A New Bio-isostere: Alkylsulphonamidophenethanolamines. *Nature* **1964,** *203,* 1283-1284.
17. Rocheblave, L.; Bihel, F.; De Michelis, C.; Priem, G.; Courcambeck, J.; Bonnet, B.; Chermann, J.; Kraus, J. Synthesis and antiviral activity of new anti-HIV amprenavir bioisosteres. *J. Med. Chem.* **2002,** *45,* 3321-3324.
18. Yoshino, K.; Kohno, T.; Morita, T.; Tsukamoto, G. Organic phosphorus compounds. 2. Synthesis and coronary vasodilator activity of (benzothiazolylbenzyl) phosphonate derivatives. *J. Med. Chem.* **1989,** *32,* 1528-1532.
19. Uno, T.; Kondo, H.; Inoue, Y.; Kawahata, Y.; Sotomura, M.; Iuchi, K.; Tsukamoto, G. Synthesis of antimicrobial agents. 3. Syntheses and antibacterial activities of 7-(4-hydroxypiperazin-1-yl)quinolones. *J. Med. Chem.* **1990,** *33,* 2929-2932.
20. Arneric, S.; Sullivan, J.; Briggs, C.; Donnelly-Roberts, D.; Anderson, D.; Raszkiewicz, J.; Hughes, M.; Cadman, E.; Adams, P.; Garvey, D.; al., e. (S)-3-methyl-5-(1-methyl-2-pyrrolidinyl) isoxazole (ABT 418): a novel cholinergic

ligand with cognition-enhancing and anxiolytic activities: I. In vitro characterization. *J. Pharmacol. Exp. Ther.* **1994,** *270,* 310-318.

21. Hynes Jr., J.; AJ, D.; Lin, S.; Wrobleski, S.; Wu, H.; Gillooly, K.; Kanner, S.; al., e. Design, Synthesis, and Anti-inflammatory Properties of Orally Active 4-(Phenylamino)-pyrrolo[2,1-f][1,2,4]triazine p38α Mitogen-Activated Protein Kinase Inhibitors. *J. Med. Chem.* **2008,** *51,* 4-16.
22. Sun, Q.; Gatto, B.; Yu, C.; Liu, A.; Liu, L.; LaVoie, E. Synthesis and evaluation of terbenzimidazoles as topoisomerase I inhibitors. *J. Med. Chem.* **1995,** *38,* 3638-3644.
23. Parks, D.; Lafrance, L.; Calvo, R.; Milkiewicz, K.; Gupta, V.; Lattanze, J.; Ramachandren, K.; Carver, T.; Petrella, E.; Cummings, M.; Maguire, D.; Grasberger, B.; Lu, T. 1,4-Benzodiazepine-2,5-diones as small molecule antagonists of the HDM2-p53 interaction: discovery and SAR. *Bioorg. Med. Chem. Lett.* **2005,** *15,* 765-770.
24. Cox, C.; Breslin, M.; Mariano, B.; Coleman, P.; Buser, C.; Walsh, E.; Hamilton, K.; Huber, H.; Kohl, N.; Torrent, M.; Yan, Y.; Kuo, L.; Hartman, G. Kinesin spindle protein (KSP) inhibitors. Part 1: The discovery of 3,5-diaryl-4,5-dihydropyrazoles as potent and selective inhibitors of the mitotic kinesin KSP. *Bioorg. Med. Chem. Lett.* **2005,** *15,* 2041-2045.
25. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1998,** *28,* 31–36.
26. Ihlenfeldt, W.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An extensible Networked Approach toward Modularity and Flexibility. *J. Chem. Inf. Comp. Sci.* **1994,** *34,* 109-116.
27. *StarDrop*, version 5.0; Optibrium: Cambridge, 2011
28. Warr, W. A. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *Comput.-Aided Mol. Des.* **2009,** *23,* 195-198.
29. Obrezanova, O.; Gola, J.; Champness, E.; Segall, M. Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des.* **2009,** *22,* 431-440.
30. Obrezanova, O.; Csanyi, G.; Gola, J.; Segall, M. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **2007,** *47,* 1847-1857.
31. Abraham, M. H.; McGowan, J. C. The use of characteristic volumes to measure cavity terms in reversed-phase liquid-chromatography. *Chromatographia* **1987,** *23,* 243-246.
32. Ertle, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000,** *43,* 3714-3717.
33. *RDKit*: Cheminformatics and Machine Learning Software. http://www.rdkit.org/ (accessed March 2, 2011).
34. Wishart, D.; Knox, C.; Guo, A.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008,** *36(Database issue),* D901-D906.
35. Perola, E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.* **2010,** *53,* 2986-2997.
36. Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry;* American Chemical Society: Washington DC, 2008; Vol. 4, pp 217-241.
37. Andrews, M.; Brown, A.; Chiva, J.; Fradet, D.; Gordon, D.; Lansdell, M.; MacKenny, M. Design and optimisation of selective serotonin re-uptake inhibitors with high synthetic accessibility: part 2. *Bioorg. Med. Chem. Lett.* **2009,** *19,* 5893-5897.
38. Glen, R. C.; A.W.R., P. A genetic algorithm for the automated generation of molecules within constraints. *J. Comput.-Aided Mol. Des.* **1995,** *9*, 181-202.

39. Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1079-1087.

40. Lameijer, E.-W.; Kok, J. N.; Back, T.; Ijzerman, A. P. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules. *J. Chem. Inf. Model.* **2006,** *46*, 545-552.

41. Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007,** *21*, 311-325.

42. Pearce, B. C.; Sofia, M. J.; Good, A. G.; Drexler, D. M.; Stock, D. A. An Empirical Process for the Design of High-Throughput Screening Deck Filters. *J. Chem. Inf. Model.* **2006,** *46*, 1060-1078.

43. Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput. Aided Mol. Des.* **2007,** *21*, 139-144.

44. Agrafiotis, D. K. Multiobjective optimization of combinatorial libraries. *IBM J. Res. Develop.* **2001,** *45*, 545-566.