Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space

Stephan C. Schürer, Prashant Tyagi, and Steven M. Muskal*

Sertanty, Inc., 9381 Judicial Drive, Suite 200, San Diego, California 92121

Received October 18, 2004

We describe a novel approach to direct the exploration of chemical space in an effort to balance synthetic accessibility and medicinal relevancy prior to experimental work. Reaction transforms containing empirical reactivity and compatibility information are dynamically assembled into reaction sequences (vProtocols) utilizing commercially available starting material feedstock. These vProtocols are evolved and optimized by a genetic algorithm, which leverages fitness functions based on predicted properties of generated molecular products. We present the underlying concepts, methodology and initial results of this prospective approach.

INTRODUCTION

Chemical synthesis is a critical element and often a limiting factor in the early drug discovery process of lead identification and optimization. Retro-synthetic analysis[1] is well established to plan a target-oriented synthesis of individual compounds or compound collections (focused libraries). However, the large amount of readily available chemical information makes it difficult to use retro-synthetic analysis efficiently and thoroughly.[2]

There have been a number of computational approaches to retro-synthetic analysis, including LHASA,[3] which applies a knowledge base of transforms and rules to identify strategic bonds. The SynGen program generates an optimal synthetic route for a target organic compound by simplification and systematization, first of skeletal dissection, then of structure and reaction characterization.[4] Gasteiger's WODCA[2,5] identifies strategic bonds based on physicochemical properties of atoms and bonds and thermodynamic stability of intermediates. SystematiChem[6] searches specific chemical reactions to arrive at synthetic routes to a target molecule. CAESA[7] estimates synthetic accessibility of target molecules based on a knowledge base of chemical reactions and available starting materials. CAMEO[8] predicts the outcome of a chemical reaction based on mechanistic reasoning. EROS[9,10] predicts the course of chemical reactions based on important electronic and energy effects in organic molecules and rules for evaluating the course of elementary processes.

Although the existing computational approaches for retro-synthetic analysis are very valuable for the analysis of individual molecules or a compound class, they are stand-alone programs and therefore not readily applicable for the high-throughput analysis of large numbers of diverse structures. Moreover, retro-synthetic analysis can only be applied with a specific idea of the target structure or target compound class, and available retro-synthesis tools are often restricted to the analysis of individual target structures.

In the early drug discovery stage of lead identification and optimization it is desirable to generate multiple ensembles of compounds that each can be accessed by a common synthetic route and allow for rapid SAR generation and optimization. Often the specific target scaffold structures may not be of primary interest—provided a relevant intellectual property position does not yet exist—and one may want to explore all target-relevant small molecule space within the scope of in-house expertise and/or otherwise accessible synthetic methodologies given readily available starting materials. From a retro-synthetic perspective it is of practical interest to quickly identify those structures and derivatives of a series of known or putative actives that can be rapidly synthesized utilizing accessible chemical methodology and available starting materials.

Here we report a novel approach to forward- and retro-synthesis based on synthetic capabilities or known chemical transformation types. We describe the directed forward synthesis of novel structures with desired (predicted) properties by assembling reaction sequences from individual chemical transformations. A genetic algorithm is utilized to dynamically combine a set of chemical transformations (transforms, described in more detail below), which utilize commercially available starting materials to yield sequences that generate likely synthesis products. Virtual synthesis protocols or vProtocols are evolved and prioritized from fitness functions based on predicted properties of generated molecular products. These vProtocols represent synthesis strategies enriched not only with the potential for synthesis success but also with the potential to produce medicinally relevant molecules.

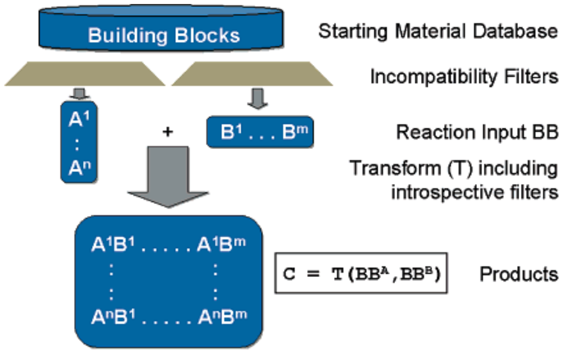
METHODOLOGY: CHEMICAL TRANSFORMS AND FILTERS

In our approach, each chemical transformation (transform) is described and loaded with information to accommodate a "mix-n-match" strategy. Here, in addition to the basic instructions represented in a chemical reaction (i.e. which atoms and bonds are rearranged, changed, added, or removed), each reaction is associated with the necessary information of chemical compatibility and/or incompatibility for each reactant along with information of chemical reactivity of the reacting functional groups, i.e., the required (or prohibited) reaction center environments. Each transform

* Corresponding author phone: (760)535-2885; fax: (775)822-1721; e-mail: smuskal@sertanty.com.

10.1021/ci0496853 CCC: $30.25
Published on Web 02/02/2005

Scheme 1. Reaction-Specific Filtering and Transformation of Building Block A and B into a Product Library C; BB Building Blocks, BB^A, BB^B Building Blocks for Reactant A, B in a Reaction $A + B \rightarrow C$



Scheme 2. Representation of the Products of a Two-Step Sequence as Transform Functions of Building Block Inputs[a]

$$A + B \longrightarrow C$$
$$C + D \longrightarrow E$$

$$\boxed{\begin{aligned} C &= T_1(BB^A,BB^B) \\ E &= T_2(C,\ BB^D) \\ &= T_2[T_1(BB^A,BB^B),BB^D] \end{aligned}}$$

[a] See Scheme 1 for explanations of symbols.

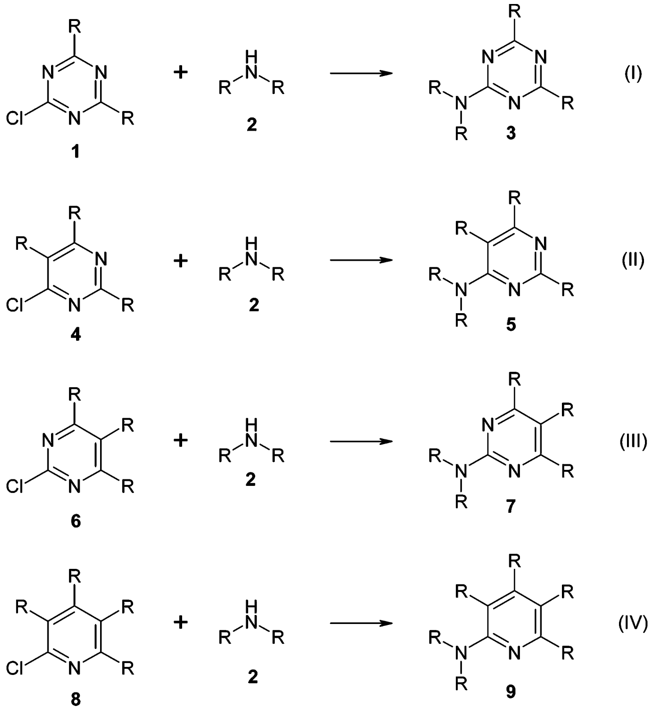
contains information to (a) select all compatible building blocks from a pool of starting materials (e.g. commercially available) that do not interfere with the reaction under the conditions based on known chemical incompatibilities, (b) select the reaction site in each building block based on chemical reactivity required for this reaction under the conditions, and (c) process instructions to generate the expected reaction products from each building block combination based on the chemical reaction mechanism while leaving the reaction center environment (e.g. transform-independent stereochemistry) in each building block unchanged (see below for more details). Scheme 1 illustrates the reaction-specific filtration of a building block database by incompatibility filters to generate compatible building block sets A and B followed by their transformation into a product library C by a transform function T. The transform function T beyond the processing instructions includes the information to select the correct reaction center environment (introspective filters) for each building block.

Using such representations, the products of a two-step reaction sequence can be represented as a transform of the products of another transform using three building block sets A, B, and D as input. Both transforms T_1 and T_2 again contain the compatibility/incompatibility and reactivity information characteristic for the reaction type they represent (Scheme 2).

As an example of these filtering concepts one can look at a nucleophilic aromatic substitution of aryl chlorides with amines (Scheme 3).

Provided similar nucleophilicy of the amine component **2** and substituents of similar electronic effect at the aryl chlorides in the reactions (I) to (IV), the reactivity toward nucleophilic aromatic substitution of the aryl chlorides is 2-chlorotriazine (**1**) > 4-chloropyrimidine (**4**) > 2-chloropyrimidine (**6**) > 2-chloropyridine (**8**). In any given reaction/reaction type we define chemical compatibility as a function of reactivity as exemplified above and exclude building blocks containing functional groups that are more reactive under the reaction conditions and therefore prone to side reactions, or which are otherwise known to interfere with the reaction unfavorably. For example in reaction (II) a 2-chloro substituent (which is less reactive that the 4-chloro) can be allowed, but in reaction (III) a more reactive 4-chloro substituent must not be present, because it would give a side reaction. There are many other general functional elements that must not be present in either of the building blocks to avoid side reactions. We define these incompatibility filters hierarchically as super- and subfilters considering reactivity types (nucleophilic, electrophilic, acidic, basic, etc.), functional group classes (acylators, alkylators, nucleophilic amines, activated aryl halides, etc.), and individual functional elements of known reactivity that can be further classified by their chemical environment. As an example, one class of basic nucleophilies includes primary or secondary alkylamines, aryl or alkyl hydrazines, unsubstituted amidine derivatives, and nucleophilic thiol compounds (all of them can be further subclassified). We leverage the powerful Daylight SMARTS language[11,12] to define such filters. In this example, the incompatibility filter can be described as shown in Figure 1.

Scheme 3. Nucleophilic Aromatic Substitution Reactions



As another example, an acylator filter includes carboxylic acid anhydrides and derivatives, acyl halides and derivatives, isocyanates or isothiocyanates and derivatives, and sulfinyl or sulfonyl halides (again with potential further subclassification) as shown in Figure 2.

Similarly, the reactivity of the amine component in reactions (I) to (IV) in Scheme 3 depends on its substituents, specifically electronic and steric effects. A primary or secondary aromatic amine is much less reactive than an aliphatic amine. Incompatibility filters are used to exclude building blocks that would lead to side reactions, i.e., if there is one or more additional equally or more nucleophilic moiety (or other reactive groups). For this it is important to express a filter of a specific functional group (substructure) in a single

```
[N;!H0;$(NC);!$([N+]);!$(NC=,#[!#6]);!$(NC=,#[#6]);!$(N[!#6]);!$(
Nc)] OR [N;!H0;$(N[N;$(N[#6]);!$(NC=,#[!#6])]);!$(NC=,#[!#6])] OR
C([NH2])=[NH] OR [S;$([SH]),$([S-])]
```

Figure 1. SMARTS representation of some basic nucleophiles defined as amine, hydrazine, amidine, or sulfide, specifically (from left to right): primary or secondary alkylamine (nitrogen with at least one hydrogen, not positively charged, no amide, thioamide or related, no enamine or ynamine, not single-bound to a non-carbon atom, i.e., no sulfonamide, etc., not bound to an aromatic carbon) OR hydrazine (no other heteroatom bound to either of the nitrogens, no hydrzide, etc.) OR amidine (nitrogen unsubstituted) OR nucleophilic sulfide (with least one hydrogen or negatively charged, includes thio acids, etc.).

```
[C;$(C(=O)OC(=O));!$(C(=O)OC(=O)[!#6]);!$(C(=O)(OC(=O))[!#6])] OR
[C;$(C(=[O,N,S])[F,Cl,Br])] OR [C;$(C(=[O,S])=N)] OR
[S;$(S(=O)[Cl,Br,F,I])]
```

Figure 2. SMARTS representation of acylators, specifically (left to right): carboxylic acid anhydride OR acyl halide or related OR isocyanate/isothiocyanate OR sulfinyl/sulfonyl halide.

```
[N;!H0;$(NC);!$([N+]);!$(NC=,#[!#6]);!$(NC=,#[#6]);!$(N[!#6]);!$(
Nc)].[N;!H0;$(NC);!$([N+]);!$(NC=,#[!#6]);!$(NC=,#[#6]);!$(N[!#6]
);!$(Nc)]
```

Figure 3. SMARTS representation of two basic primary or secondary alkylamines (two dot-separated primary or secondary alkylamines as described in component 1 of Figure 1).

```
[c;$(c1(Cl)nc(nc(c1[A,a])[A,a])[A,a]):12]1(Cl)[a:11]([a:10]([a:7]
[a:9]([A,a:3])[a:8]1)[A,a:4])[A,a:5].[N;$([N;!H0]([A,a])[A,a]);$(
NC);!$([N+]);!$(NC=,#[!#6]);!$(NC=,#[#6]);!$(N[!#6;!#1]);!$(Nc):6
]([A,a:1])([A,a:2])[H]>>[c:12]1([N:6]([A,a:1])[A,a:2])[a:8][a:9](
[A,a:3])[a:7][a:10]([A,a:4])[a:11]1[A,a:5]
```

Figure 4. SMIRKS representation of reaction (II) in Scheme 3; reactant component 1 represents the 4-chloropyrimidine component with the 4-carbon as the reaction center atom, reactant 2 is an primary or secondary alkylamine as the one shown in Figure 1 with nitrogen as the reaction center atom; nonreaction center atoms are generalized as [A,a] strings to transform their local environments and atom properties independent from these environments/properties (see text).

string so double occurrences can easily be defined and queried as SMARTS matches. As an example, two or more primary or secondary alkylamines are defined as shown in Figure 3.

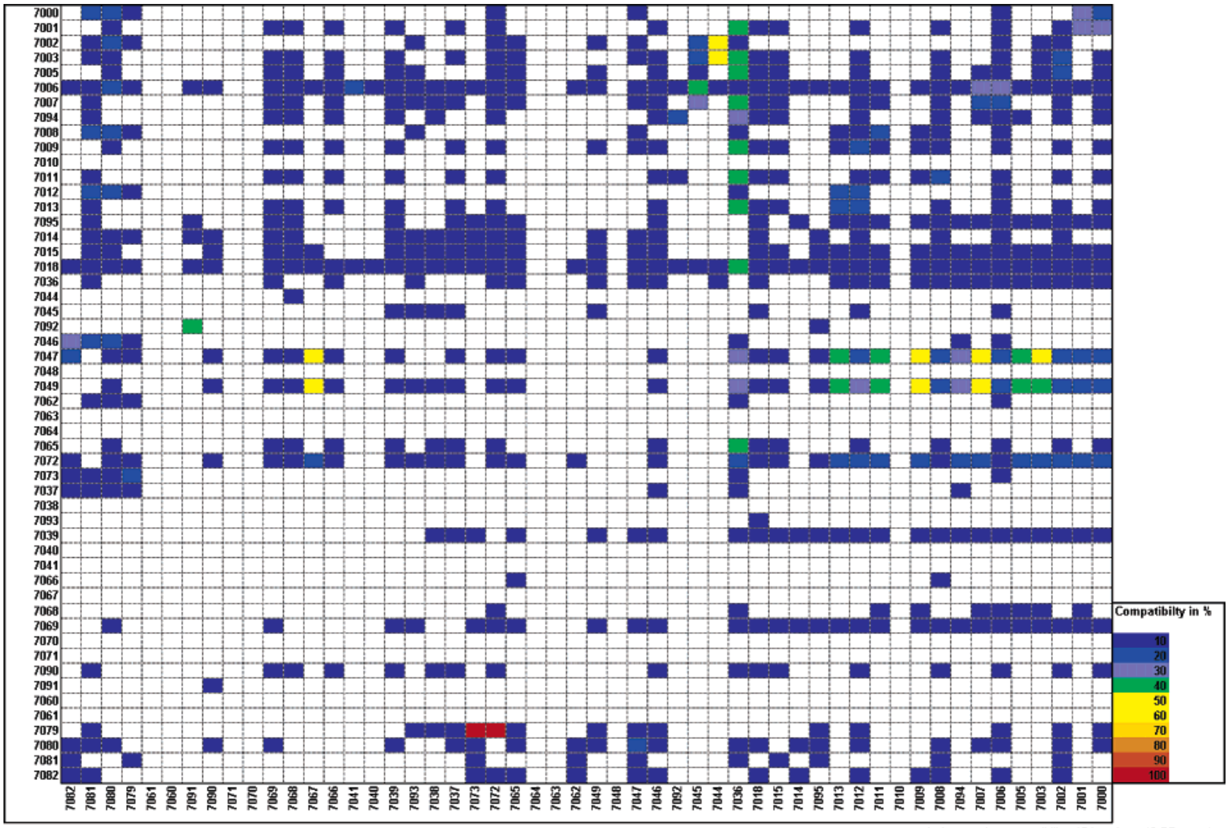
A definition of a double occurrence of a primary or secondary alkylamine using an OR combination of e.g. primary alkyl- OR secondary alkylamine would be more difficult to process.

In addition to the definition of incompatibility of certain more or equally reactive (or otherwise incompatible) functional elements with a reaction transformation, often a specific reaction site (reaction center) must be chosen by the transform function. Generally a less reactive functional group must be allowed in any building block, e.g. an arylamine in the presence of an alkylamine. For example, if in reaction (II) the amine component is defined as a primary or secondary alkylamine, this information must be incorporated into the reaction transform in order to specifically select this nucleophilic amine vs any other potential amine or amide that could match the generic representation of the reaction. Again, the SMIRKS language[11,13] can be used for the definition of these reaction transforms, provided we semantically and syntactically separate the reaction center atoms from their environment. This has a number of advantages. Generally such transforms—generated from common reaction drawings—more closely transform the reactants according to the chemical reaction mechanism disregarding restrictions that may be intrinsic in the syntax of the original reaction representation. Furthermore, this approach allows incorporation of additional reactivity requirements into the reaction transforms, i.e., a more specific description of the required reaction center environment. For example, we can again look at reaction (II) as expressed in the SMIRKS transform shown in Figure 4.

The reaction center atoms are explicitly defined in the generic reactants and products, and their environments are defined as recursive components only in the reactants to ensure a unique match.[13] All nonreaction center atoms are generalized, and their local environments and atom properties (like chirality, isotopic information, substitution pattern, etc.) are transformed into the products independent from those properties themselves.

The amine reactant (second component in the SMIRKS) shows a number of restrictions all referring to the amine nitrogen reaction center atom including the following: at least one hydrogen substituent, bound to an aliphatic carbon, not positively charged, not bound to a carbon with a double or triple bond to a non-carbon atom (no amide, amidine, etc.), not bound to a carbon with a double or triple bond to a carbon atom (no enamine or ynamine), not bound to any atom that is not carbon or hydrogen (no hydrazine or hydroxylamine derivative, sulfonamide or derivative, etc.), and not bound to an aromatic carbon.

To build these introspective filters it is important to define required reactivity of a chemical substructure (e.g. functional group) such that all information refers to a single atom that



pairwise reaction compatibility, 15 iterations, 10 BB samples

Figure 5. Heat map of pairwise reaction compatibility of 2-step reaction sequences in %; 10 samples, 15 iterations. While compatibility is based on successful processing of molecules through pairwise reaction combinations, the assessment of reaction compatibility can be further enriched with experimental synthetic yields. Compatibility is not symmetric. Data in Figure 5 are shown as compatibility of reactions shown as *y*-axis (vertical at the left side of the figure) as a function of reactions on *x*-axis (horizontal at the bottom of the figure); products of reactions on *x*-axis feed into reactants of reactions on *y*-axis.

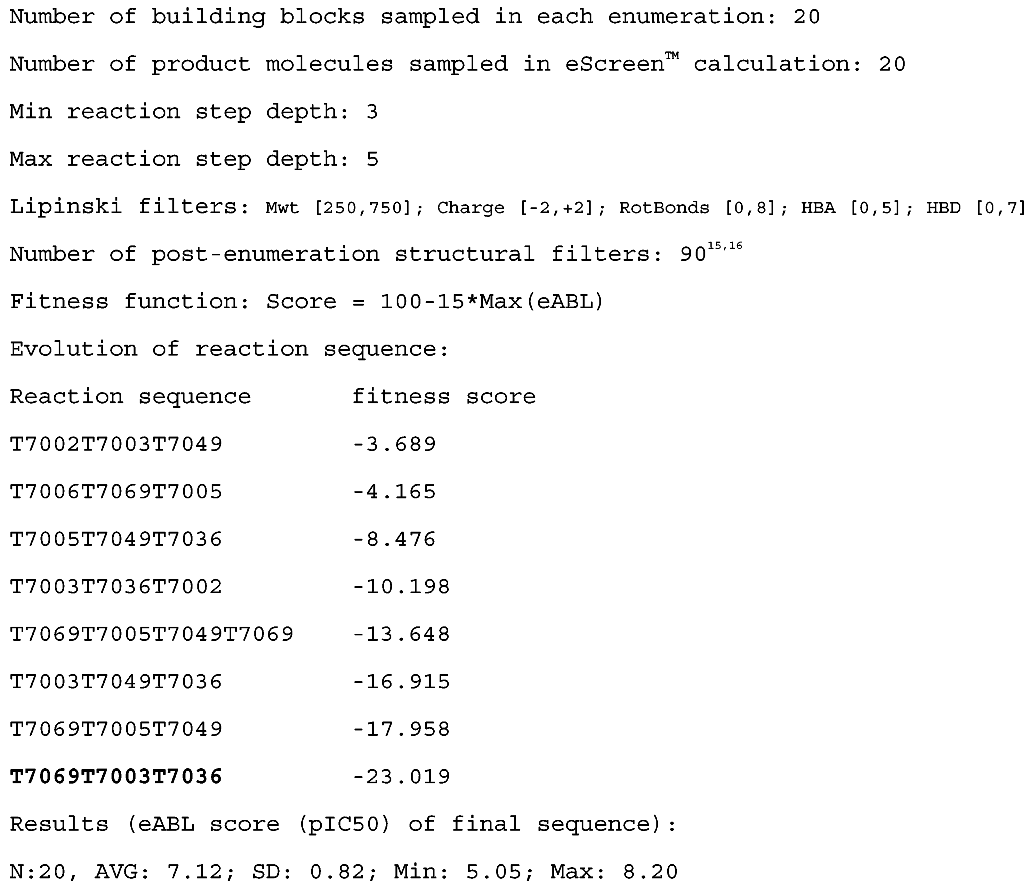
corresponds to a reaction center atom (recursive elements in the SMARTS language facilitate this). Similarly, these filters can be classified in super- and subcategories as long as they refer to the same functional group atom (reaction center atom). As examples, the filters components shown above in Figures 1 to 4 each match a single atom within a substructure defining the chemical environment that influences chemical reactivity.

We built a user interface, which facilitates the nonspecialist definition of such filters in a hierarchical way. This interface allows the build-up of super- and subfilters and facilitates the association of incompatibility and introspective filters with reaction transformations, either as exclusion filters or as required functionalities.

Given a large structure−activity knowledgebase for kinase inhibitors,[14] we chose a few reaction types relevant for the synthesis and modification of heterocyclic kinase inhibitors and defined a number of reaction representations for each of these classes. We defined transforms for nucleophilic aromatic substitutions of activated aryl halides by alkyl and arylamines including chlorotriazine, chloropyrimidine, chloro- or fluoropyridine derivatives, chloro- or fluoronitrobenzene derivatives; amine acylation reactions, including formation of carboxylic acid amides, carbamates, ureas, thioureas; syntheses of heterocyclic scaffolds including annelated aryl pyrimidinones, pyrimidindiones, oxazoles, thiazoles, and imidazoles, most importantly quinazolinone-, quinazolindione-, benzoxazole-, benzothiazole-, and benzimidazole derivatives; generalized representations for Pd-catalyzed arylamination and Suzuki coupling reactions. We also included simple functional group transformations to utilize masked functionalities such as aryl nitro reduction to give arylamines, transformation of pyrimidinones and -diones and related compounds to give the respective chloropyrimidine and related compounds, and standard deprotection steps.

For this relatively small reaction basis-set, we defined the most important and common functional groups—the filters as described above—which we considered most relevant to describe chemical compatibility and reactivity and associated the respective incompatibility and introspective filters with each individual reaction transform.

Beyond the filters associated with the reactions we defined a set of global filters of undesired fragments to be excluded in any final product and/or reactant to ensure the generation of medicinally more relevant products. Such filters include >50 representations of reactive functionalities and substructures including alkylators and acylators, electrophiles and certain nucleophiles (like thiols or hydrazines)[15] and >35 representations of other undesired motifs that are likely to result in medicinally unfavorable properties including all structures that do contain any nonstandard element, structures with too many aryl halides or aryl nitro groups, extended conjugated systems, crown ethers, fluororganic compounds, etc.[16] In addition we applied adjusted Lipinski constraints[17,18] using slightly improved SMARTS queries for hydrogen bond donors, acceptors, and rotatable bonds as compared to commonly used definitions.[12]



```
Number of building blocks sampled in each enumeration: 20

Number of product molecules sampled in eScreen™ calculation: 20

Min reaction step depth: 3

Max reaction step depth: 5

Lipinski filters: Mwt [250,750]; Charge [-2,+2]; RotBonds [0,8]; HBA [0,5]; HBD [0,7]

Number of post-enumeration structural filters: 90[15,16]

Fitness function: Score = 100-15*Max(eABL)

Evolution of reaction sequence:

Reaction sequence       fitness score

T7002T7003T7049         -3.689

T7006T7069T7005         -4.165

T7005T7049T7036         -8.476

T7003T7036T7002         -10.198

T7069T7005T7049T7069    -13.648

T7003T7049T7036         -16.915

T7069T7005T7049         -17.958

T7069T7003T7036         -23.019

Results (eABL score (pIC50) of final sequence):

N:20, AVG: 7.12; SD: 0.82; Min: 5.05; Max: 8.20
```

Figure 6. Parameters and evolution for a GA simulation against an ABL-kinase eScreen model (eABL).

METHODOLOGY: SURVEY OF CHEMICAL SPACE AND GENETIC ALGORITHM

With the above introduced definition of the products E of a two-step reaction as two transform functions of building blocks (see Scheme 2) as $E = T_2[T_1(BB^A,BB^B),BB^D]$ one can now describe a subset of these products by applying additional filter functions, such as Lipinski constraints F^{Lip} or undesired fragments $F^{BadFrag}$ as in formula (I) below.

$$E^{Subset} = F^{BadFrag}[F^{Lip}(E)] = F^{BadFrag}[F^{Lip}\{T_2[T_1(BB^A,BB^B),BB^D]\}] \quad (I)$$

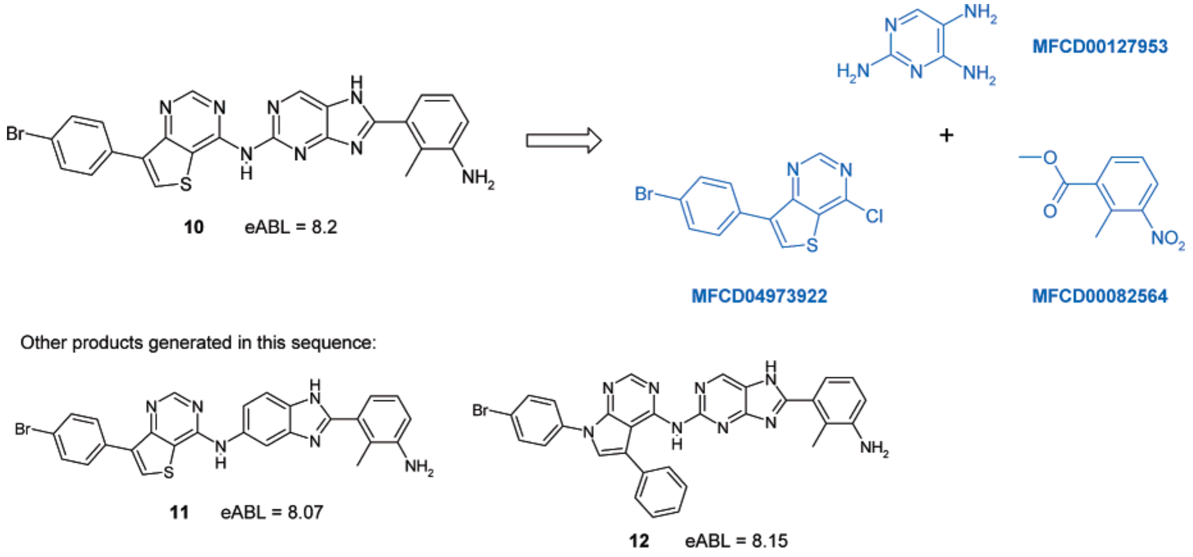
In addition to these descriptor and substructure filters we applied the Sertanty eScreen (QSAR) technology[19] to prioritize compounds with likely kinase activity. These eScreen models are based on a large number of structure–activity data points[14] using 3D pharmacophoric fingerprints[20,21] and have proven significant enrichment capability in compound prioritization/selection efforts. A similar methodology was recently independently applied to protein kinases.[22]

With the definitions of the reaction transforms, extensive substructure and Lipinski-type filters and QSAR characterization, it was our goal to explore the chemical space defined by the basis-set chemistries starting from commercially available building blocks considering novelty, drug-likeness, and predicted activity/selectivity against kinase targets.

Our initial attempt was to systematically build all possible reaction sequences from these basis-set reaction transforms and generate the products of these sequences using building blocks obtained by filtering ACD[23] (applying building block compatibility filters associated with the transforms). Although this approach generated a large number of novel structures after Lipinski filtering and elimination of undesired structural motifs,[24] we realized that it would be unfeasible to systematically sample the chemical space defined by even a small number of reactions with reasonable computational effort.

As an alternative to systematic exploration of this chemical transform space, we looked for an approach to identify the best sequences of chemical reactions leading to product structures with desired (predicted) properties starting from commercially available starting materials. Some reaction sequences will naturally generate a larger number of final products than others; e.g. aniline products obtained by reduction of aryl nitro derivative will often be compatible with transformations requiring arylamines as reactants. Such sequences are likely to generate more products than reaction sequences that require building blocks with orthogonal functionalities (functionalities that independently react in different chemical transformations under different reaction conditions). As a first step to quantitatively analyze such preferred reaction compatibility based on commercially available starting materials, we generated a matrix of pairwise transform compatibility scores (Figure 5). The results are represented as a heat map in % compatibility as ratios of successfully generated final products to expected products (i.e. building block input obtained as products of reaction step 1). Sampling was performed in 15 iterations with 10 randomly selected building block combinations each. The identifiers for the individual reactions in Figure 5 and the

Scheme 4. Selected Products with Starting Materials of GA-Simulation Outlined in Figure 6



reaction representations are provided in the Supporting Information.

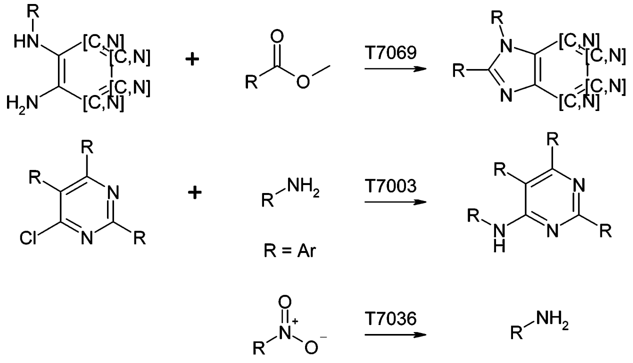
To survey the chemical space defined by our reaction transforms more efficiently with the goal of identifying the best reaction sequences generating products with desired properties, we applied a genetic algorithm (GA) to more efficiently direct the exploration of synthetic possibility.[25−28]

The GA uses a text-based representation of reaction step transforms and sequences (see Scheme 2) and starts with a random population of reaction sequences of one to a defined maximum number of reaction steps. Reaction sequences are evaluated upon enumerating all products applying the respective transform functions starting from prefiltered building blocks as described above.

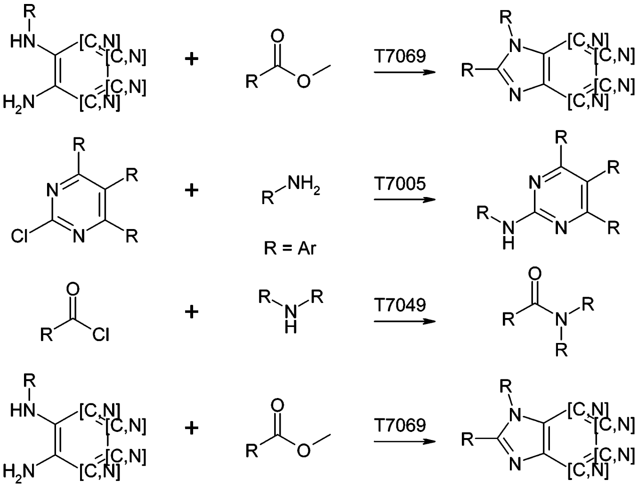
Generated products are filtered by Lipinski descriptors and then by structural filters. Remaining products are scored by predicted activity using quantitative kinase target-specific or binary kinase ATP binding eScreen models.[19] In practice, any computational model which accepts standardized molecule file formats (e.g. MDL's SDFile format, Daylight SMILES/TDTFile formats, etc.) and produces a numerical assessment of the molecules in a readily digestible format can be leveraged in the GA-fitness function. Reaction sequences have been optimized by crossover mutation of the highest scoring sequences to maximize fitness. Throughout the simulation the top performing sequences and their generated products are captured. A detailed algorithmic outline of the genetic algorithm is provided in the Supporting Information. For the generation of reaction sequence populations, multiple copies of the reaction transforms are offered and—based on precomputed pairwise reaction compatibility (see Figure 5)—highly effective two-step sequences are additionally introduced as starting 'genes'.

As an example of an initial simulation we describe the evolution of a three-step sequence optimized against our ABL-kinase (Abelson murine leukemia viral oncogene homolog) eScreen model using the simulation parameters given in Figure 6.

Scheme 5. GA-Generated Reaction Sequence To Produce Products in Scheme 4



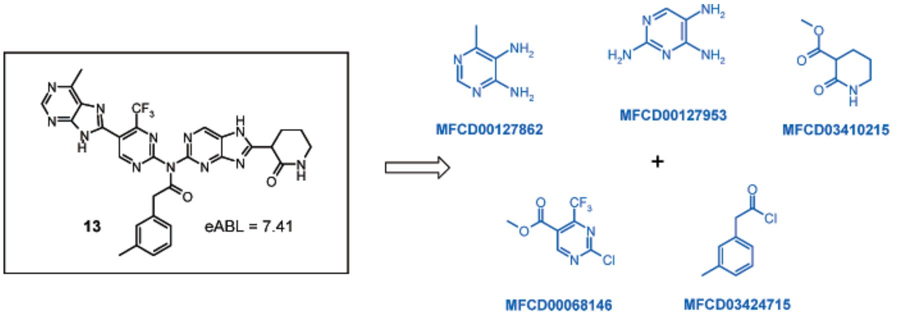
Scheme 6. GA-Generated Reaction Sequence T7069T7005T7049T7069



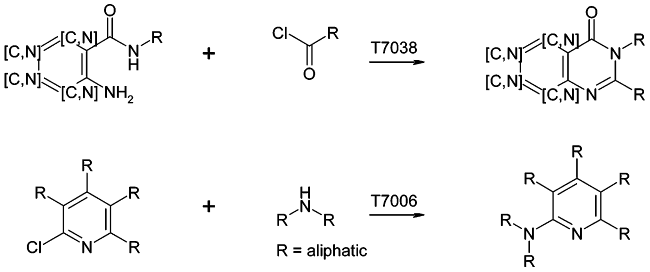
SIMULATION RESULTS AND DISCUSSION

As shown in Figure 6 each reaction sequence generation produced products with a lower scoring value corresponding to increased predicted eABL activity. Scheme 4 highlights some products (**10**, **11**, **12**) and the respective eABL scores of the GA-generated final reaction sequence T7069T7003-T7036. Scheme 5 shows the reaction sequence generated by the GA leading to these structures. Reactions 7069 to 7003 and 7003 to 7036 show relatively low compatibility in Figure 5, because only a small subset of the building blocks compatible with the first reaction will also have a functionality to react in the second transformation, etc. Nonetheless the obtained reaction sequence emerged to the top, because

Scheme 7. GA-Generated Product 13 and Starting Materials for Sequence T7069T7005T7049T7069



Scheme 8. GA-Generated vProtocol T7038T7006



some prospective active compounds are generated. While the GA optimizes for a reaction sequence, thus indirectly also a subset of building blocks—with required orthogonal functionalities—are selected. The aniline functionality in the simulation products could serve as a site for further derivatization, and similar the aryl bromide for some of the products.

Scheme 6 shows a reaction sequence that emerged as result of a similar simulation using the same parameters as before. The average eABL score for compounds generated in this vProtocol is 6.58 with SD = 0.67 (N = 20), eABL^{min} = 5.37 and eABL^{max} = 7.58. Whereas transformation 7069 and 7005 only show low compatibility, 7005 and 7049 is much more compatible, because the 7005 secondary amine product can be acylated by 7049 in many cases.

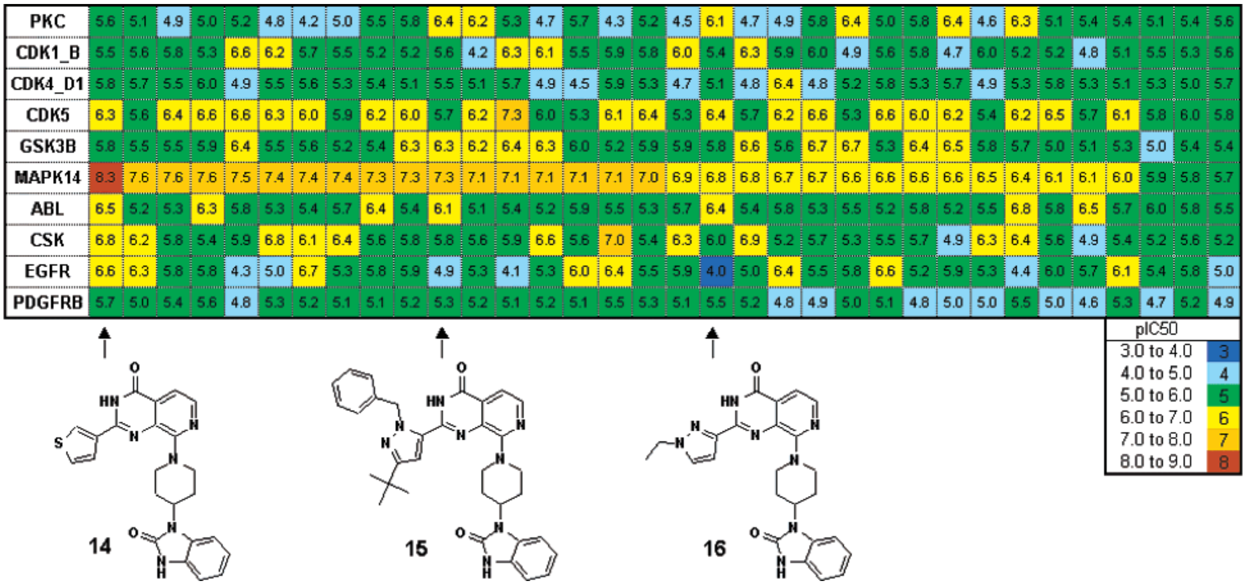
An example product and its commercially available starting materials are shown in Scheme 7.

These results exemplify the concept of the directed evolution of reaction sequences toward protocols that generate prospective active compounds using a quantitative eScreen model.

For subsequent simulations we incorporated a binary ATP-binding site classification pharmacophore-based model[29] into the scoring function of the GA to explore more general ATP-binding site directed kinase inhibitors. We also allowed simple two-step sequences. For the highest scoring vProtocols the quantitative eScreen activities were calculated for a few kinase targets representing members of different kinase groups and families: PKC, CDK1, CDK4, CDK5, GSK3B, MAPK14, ABL, CSK, EGFR, PDGFR. As a result of such a simulation a generated 2-step protocol is shown in Scheme 8.

The eSceen scores as pIC50 values of 34 generated compounds and selected structures (**14**, **15**, **16**) are shown in Figure 7. Some of the compounds display predicted activity for p38-alpha (MAPK14).

Although obtained results of our initial simulations demonstrate the concept of the GA-enabled directed evolution of reaction sequences generating products with desired (prospective) properties from commercially available starting materials, the molecular structures obtained have a rather high molecular weight. We therefore reduced the maximum molecular weight to 450 in order to produce more favorable structures; all other filters were unmodified. With reduced molecular weight a similar vProtocol to T7038T7006 shown in Scheme 8 emerged as high-scoring in the genetic algorithm. Some of the results are shown in Table 1 below (structures **17** to **20**).



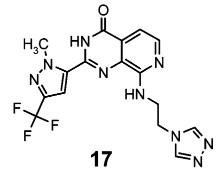
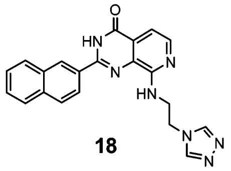
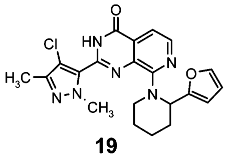
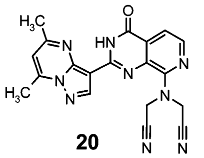
PKC	5.6	5.1	4.9	5.0	5.2	4.8	4.2	5.0	5.5	5.8	6.4	6.2	5.3	4.7	5.7	4.3	5.2	4.5	6.1	4.7	4.9	5.8	6.4	5.0	5.8	6.4	4.6	6.3	5.1	5.4	5.4	5.1	5.4	5.6
CDK1_B	5.5	5.6	5.8	5.3	6.6	6.2	5.7	5.5	5.2	5.2	5.6	4.2	6.3	6.1	5.5	5.9	5.8	6.0	5.4	6.3	5.9	6.0	4.9	5.6	5.8	4.7	6.0	5.2	5.2	4.8	5.1	5.5	5.3	5.6
CDK4_D1	5.8	5.7	5.5	6.0	4.9	5.5	5.6	5.3	5.4	5.1	5.5	5.1	5.7	4.9	4.5	5.9	5.3	4.7	5.1	4.8	6.4	4.8	5.2	5.8	5.3	5.7	4.9	5.3	5.8	5.3	5.1	5.3	5.0	5.7
CDK5	6.3	5.6	6.4	6.6	6.6	6.3	6.0	5.9	6.2	6.0	5.7	6.2	7.3	6.0	5.3	6.1	6.4	5.3	6.4	5.7	6.2	6.6	5.3	6.6	6.0	6.2	5.4	6.2	6.5	5.7	6.1	5.8	6.0	5.8
GSK3B	5.8	5.5	5.5	5.9	6.4	5.5	5.6	5.2	5.4	6.3	6.3	6.2	6.4	6.3	6.0	5.2	5.9	5.9	5.8	6.6	5.6	6.7	6.7	5.3	6.4	6.5	5.8	5.7	5.0	5.1	5.3	5.0	5.4	5.4
MAPK14	8.3	7.6	7.6	7.6	7.5	7.4	7.4	7.4	7.3	7.3	7.3	7.1	7.1	7.1	7.1	7.1	7.0	6.9	6.8	6.8	6.7	6.7	6.6	6.6	6.6	6.6	6.5	6.4	6.1	6.1	6.0	5.9	5.8	5.7
ABL	6.5	5.2	5.3	6.3	5.8	5.3	5.4	5.7	6.4	5.4	6.1	5.1	5.4	5.2	5.9	5.5	5.3	5.7	6.4	5.4	5.8	5.3	5.5	5.2	5.8	5.2	5.5	6.8	5.8	6.5	5.7	6.0	5.8	5.5
CSK	6.8	6.2	5.8	5.4	5.9	6.8	6.1	6.4	5.6	5.8	5.8	5.6	5.9	6.6	5.6	7.0	5.4	6.3	6.0	6.9	5.2	5.7	5.3	5.5	5.7	4.9	6.3	6.4	5.6	4.9	5.4	5.2	5.6	5.2
EGFR	6.6	6.3	5.8	5.8	4.3	5.0	6.7	5.3	5.8	5.9	4.9	5.3	4.1	5.3	6.0	6.4	5.5	5.9	4.0	5.0	6.4	5.5	5.8	6.6	5.2	5.9	5.3	4.4	6.0	5.7	6.1	5.4	5.8	5.0
PDGFRB	5.7	5.0	5.4	5.6	4.8	5.3	5.2	5.1	5.1	5.2	5.3	5.2	5.1	5.2	5.1	5.5	5.3	5.1	5.5	5.2	4.8	4.9	5.0	5.1	4.8	5.0	5.0	5.5	5.0	4.6	5.3	4.7	5.2	4.9

pIC50	
3.0 to 4.0	3
4.0 to 5.0	4
5.0 to 6.0	5
6.0 to 7.0	6
7.0 to 8.0	7
8.0 to 9.0	8

14 15 16

Figure 7. eScreen scores and structures generated in vProtocol T7038T7006.

Table 1. Generated Products and Predicted eScreen Activity

Structure	PKC	CDK1	CDK4	CDK5	GSK3B	MAPK14	ABL	CSK	EGFR	PDGFRB
 **17**	5.8	4.7	5.0	6.2	6.0	5.1	5.5	6.6	5.4	5.2
 **18**	5.2	5.5	5.5	5.4	6.0	6.2	5.6	5.5	6.2	5.4
 **19**	5.0	5.7	5.9	6.5	5.5	5.1	5.0	5.8	4.9	4.9
 **20**	5.8	4.7	5.3	6.2	5.3	5.1	5.1	6.3	4.6	5.3

For these results it should be noted that although up to 5-step sequences were considered by the GA, 2-step sequences emerged due to requirements of the final products (Lipinski and structural constraints) before evaluation by the binary classification-based scoring function of the GA.

Based on the simulation results exemplified in Figure 7 and Table 1—specifically the suggested MAKP14 activity of compounds **14**, **15**, **16**, **18**—we calculated Tanimoto similarity values of simulation results with known kinase inhibitor from our database[14] using Daylight fingerprints[30] for structural similarity and pharmprint fingerprints[21,22] for pharmacophoric similarity. Similarity values of structure **15** to three known MAPK14 inhibitors **21**, **22**, **23** are shown in Figure 8.

21 is an inhibitor developed by Bayer,[31] **22** is the well-known Boehringer Ingelheim kinase inhibitor BIRB 796,[32,33] and **23** has been developed by Astra Zeneca.[34] Clearly similarity values confirm the novelty of **15** compared to **21**, **22**, **23**. Where there is some degree of structural similarity of **15** and **21**, pharmacophoric similarity is very low. Diarylurea kinase inhibitors such as **21** have been shown to interact with an allosteric binding site of MAPK14, which

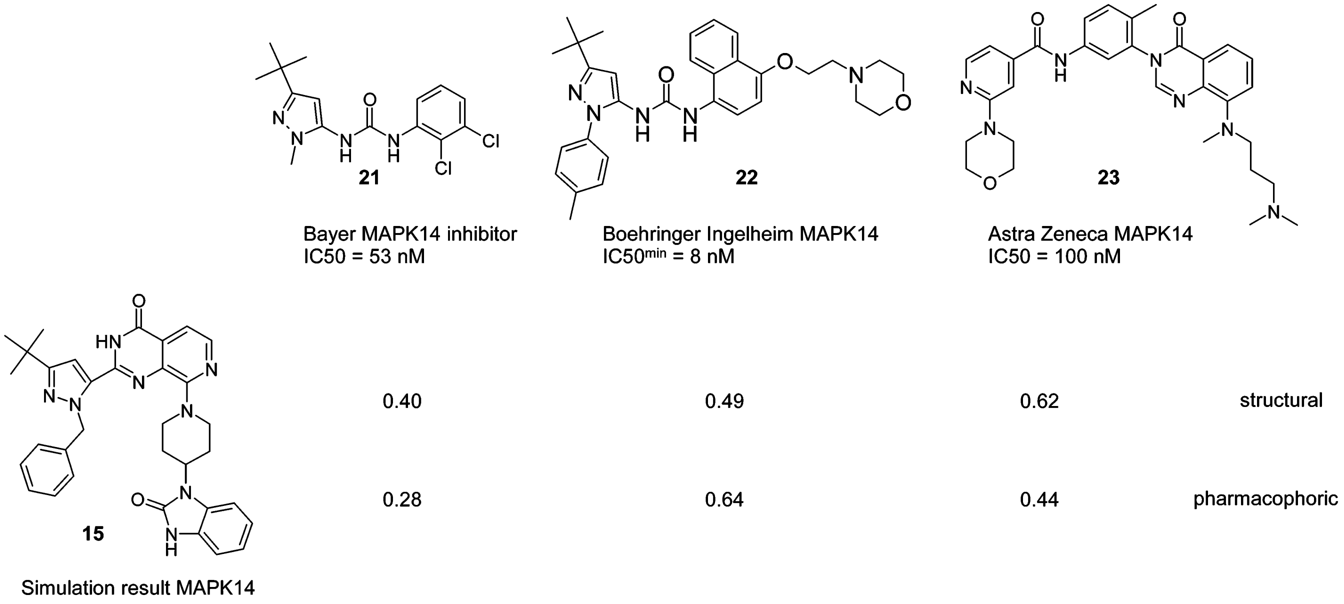
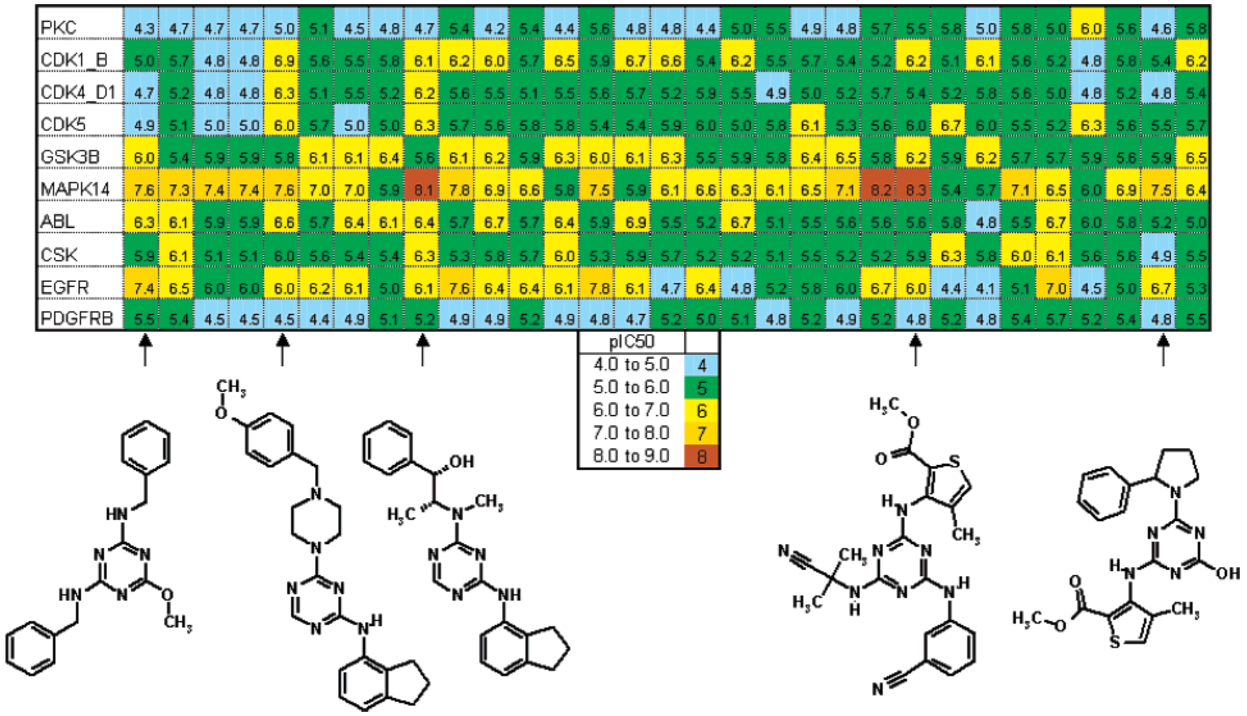


Figure 8. Structural and pharmacophoric similarity of structure **15** to known MAPK14 kinase inhibitors **21**, **22**, **23**.

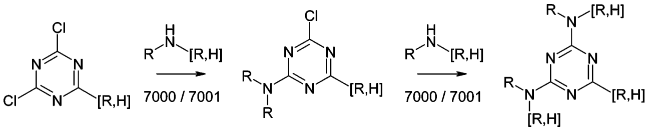


PKC	4.3	4.7	4.7	4.7	5.0	5.1	4.5	4.8	4.7	5.4	4.2	5.4	4.4	5.6	4.8	4.8	4.4	5.0	5.5	4.9	4.8	5.7	5.5	5.8	5.0	5.8	5.0	6.0	5.6	4.6	5.8
CDK1_B	5.0	5.7	4.8	4.8	6.9	5.6	5.5	5.8	6.1	6.2	6.0	5.7	6.5	5.9	6.7	6.6	5.4	6.2	5.5	5.7	5.4	5.2	6.2	5.1	6.1	5.6	5.2	4.8	5.8	5.4	6.2
CDK4_D1	4.7	5.2	4.8	4.8	6.3	5.1	5.5	5.2	6.2	5.6	5.5	5.1	5.5	5.6	5.7	5.2	5.9	5.5	4.9	5.0	5.2	5.7	5.4	5.2	5.8	5.6	5.0	4.8	5.2	4.8	5.4
CDK5	4.9	5.1	5.0	5.0	6.0	5.7	5.0	5.0	6.3	5.7	5.6	5.8	5.8	5.4	5.4	5.9	6.0	5.0	5.8	6.1	5.3	5.6	6.0	6.7	6.0	5.5	5.2	6.3	5.6	5.5	5.7
GSK3B	6.0	5.4	5.9	5.9	5.8	6.1	6.1	6.4	5.6	6.1	6.2	5.9	6.3	6.0	6.1	6.3	5.5	5.9	5.8	6.4	6.5	5.8	6.2	5.9	6.2	5.7	5.7	5.9	5.6	5.9	6.5
MAPK14	7.6	7.3	7.4	7.4	7.6	7.0	7.0	5.9	8.1	7.8	6.9	6.6	5.8	7.5	5.9	6.1	6.6	6.3	6.1	6.5	7.1	8.2	8.3	5.4	5.7	7.1	6.5	6.0	6.9	7.5	6.4
ABL	6.3	6.1	5.9	5.9	6.6	5.7	6.4	6.1	6.4	5.7	6.7	5.7	6.4	5.9	6.9	5.5	5.2	6.7	5.1	5.5	5.6	5.6	5.6	5.8	4.8	5.5	6.7	6.0	5.8	5.2	5.0
CSK	5.9	6.1	5.1	5.1	6.0	5.6	5.4	5.4	6.3	5.3	5.8	5.7	6.0	5.3	5.9	5.7	5.2	5.2	5.1	5.5	5.2	5.2	5.9	6.3	5.8	6.0	6.1	5.6	5.6	4.9	5.5
EGFR	7.4	6.5	6.0	6.0	6.0	6.2	6.1	5.0	6.1	7.6	6.4	6.4	6.1	7.8	6.1	4.7	6.4	4.8	5.2	5.8	6.0	6.7	6.0	4.4	4.1	5.1	7.0	4.5	5.0	6.7	5.3
PDGFRB	5.5	5.4	4.5	4.5	4.5	4.4	4.9	5.1	5.2	4.9	4.9	5.2	4.9	4.8	4.7	5.2	5.0	5.1	4.8	5.2	4.9	5.2	4.8	5.2	4.8	5.4	5.7	5.2	5.4	4.8	5.5

pIC50	
4.0 to 5.0	4
5.0 to 6.0	5
6.0 to 7.0	6
7.0 to 8.0	7
8.0 to 9.0	8

Figure 9. eScreen scores and structures generated in vProtocol T7000/7001T7000/7001.

Scheme 9. GA-Generated vProtocol T7000/7001T7000/7001



is spatially distinct from the ATP binding pocket.[32] Pharmacophoric dissimilarity of **15** and **21** is explained by the ATP-binding-directed simulation that produced the structures in Figure 7 and Table 1. **22** was developed from SAR of similar diaryl urea structures toward establishing additional binding interactions with the ATP pocket[32,33] and therefore may be more similar with respect to pharmacophores. Based on the structure of **23** and its activity in an ATP-competitive assay we assume that it binds at the ATP pocket.[34]

In another simulation with similar parameters, we obtained sequences of two subsequent nucleophilic aromatic substitutions of 2,4-dichlorotriazine derivatives with alkyl or arylamines to give the respective 2,4-diaminotriazine derivatives (Scheme 9).

Some example diaminotriazines and their calculated activities against the same targets are shown in Figure 9.

Triazines are a well-known class of compounds, and some of the obtained structures show structural and pharmacophoric similarity (range 0.4 to 0.7) to reported triazine kinase inhibitors.[35]

The above results demonstrate that meaningful structures can be obtained using our ATP-binding site classification as the guiding function in GA simulations. More importantly novel structures are obtained via accessible (predefined) reaction chemistries from commercially available starting materials. It is important to keep in mind that eScreen-predicted activities should be seen as a statistical enrichment and will not always be accurate for individual compounds. Enrichment studies for several of our models including MAPK14 are provided in reference 19.

Besides the forward exploration of chemical space, a similar methodology can be applied for retro-synthetic analysis using inverse or retro-transform functions. Commercial availability in conjunction with price and number of reaction steps can be incorporated in the scoring function of the GA. This would allow the exploration of the most efficient synthetic sequences leading to ensembles of provided (not just individual) compounds. We are currently exploring the application of reverse-transforms for retro-synthetic analysis for compound collections, which will be reported in due course.

SUMMARY AND CONCLUSION

We present a novel approach for the exploration of synthetically feasible small molecule chemical space from commercially available starting materials, directed toward medicinally relevancy, applying predictive computational QSAR models and a number of physicochemical and structural filters. We have developed transform functions that facilitate synthetically meaningful processing of chemical reactions incorporating information of chemical compatibility and reactivity. A genetic algorithm was applied to survey reaction sequences assembled from such transforms using predicted properties of the generated final products as a feedback function. As initial results, we presented and discussed generated reaction sequences, product structures and predicted properties for a couple of simulations. We are currently expanding the set of transforms and will report results of future simulations as well as retro-synthetic applications.

ACKNOWLEDGMENT

This work was financed in part by the National Institute of Standards and Technology (NIST) in its ATP program: ‘Chemical Intelligence Platform for Rapid Discovery of Drug Leads’ (grant No. 2002005127).

Supporting Information Available: Tables of the chemical reaction representations applied in the simulations with brief descriptions of compatibility and flowcharts describing the genetic algorithm. This material is available free of charge at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Corey, E. J.; Chen, X.-M. *The Logic of Chemical Synthesis*; John Wiley & Sons: New York, 1989.

(2) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613−2633.

(3) Selected reference: Computer-Assisted Design of Complex Organic Syntheses. Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178−192; also see http://lhasa.harvard.edu.

(4) "Generating Benign Alternative Syntheses − the SynGen Program", Henrickson, J. B. *ACS Symposium Series #823*, Chapter 10, 127−144, 2002; also see http://syngen2.chem.brandeis.edu.

(5) "The WODCA System" Gasteiger, J.; Ihlenfeldt, W. D. In *Software-Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1990, 57−65; also see http://www2.chemie.uni-erlangen.de/software/wodca/index5.html

(6) SystematiChem, SysChem Inc, Coeur d'Alene ID, USA.; http://www.syschem.com/.

(7) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.* **1995**, *3*, 34−50; also see http://www.simbiosys.ca/caesa/index.html.

(8) Selected reference: Salatin, T. D.; Jorgensen, W. J. Computer-assisted mechanistic evaluation of organic reactions. 1. Overview. *J. Org. Chem.* **1980** *45*, 2043−2050; also see http://zarbi.chem.yale.edu/products/cameo/index.shtml.

(9) Höllering, R.; Gasteiger, J.; Steinhauer, L.; Schulz, K.-P.; Herwig, A. Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482−494.

(10) EROS − A Computer Program for Generating Sequences of Reactions Gasteiger, J.; Jochum, C. *Topics Curr. Chem.* **1978**, *74*, 93−126; also see http://zabib.chemie.uni-erlangen.de/software/eros/index.html.

(11) *Daylight Theory Manual,* Chapter 4, Daylight Chemical Information Systems, Santa Fe, NM, USA; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

(12) For example SMARTS see Daylight Chemical Information Systems: http://www.daylight.com/support/faq/smarts_examples.faq.html.

(13) *Daylight Theory Manual,* Chapter 7, Daylight Chemical Information Systems, Santa Fe, NM, USA; http://www.daylight.com/dayhtml/doc/theory/theory.rxn.html.

(14) Sertanty, Inc., San Diego, CA. Kinase knowledge base of > 100,000 data points and > 300,000 unique kinase inhibitor molecules; http://www.sertanty.com/prod/content/kinase.html.

(15) Specifically these reactive functionalities and substructures were excluded: alkyl halides; alkyl, acyl sulfonates, sulfates, etc.; trichloroacetimidate and related derivatives; epoxides, aziridines, etc.; pyrilium salts and related; nitrogen or oxygen halides; sulfur halogen; allenes; ketenes; azo, diazo compounds, azides; nitroso compounds; acyl halides and related; anhydrides, bicarbonates and related; oxysuccinimides and derivatives; isocyanates, carbodiimides and related; isonitriles; activated esters and derivatives; ortho esters (acyclic); sulfinyl, sulfonyl halides, etc.; Michael acceptors; peroxides, disulfides and related; aldehydes, thioaldehydes and related; Schiff bases, imine derivatives; thioesters, thio acids and related; thiols, thiolates; hydrazines; hydroxylamines; sulfonic, sulfinic acids; reactive halo-aryl derivatives.

(16) Specifically these undesired fragments were excluded: any element except C, H, N, O, S, B, F, Cl, Br, I.; nonstandard isotopes of these elements; extended aromatic carbon systems; >=4 halides; fluoro-organic compounds; unbranched chains >=5 atoms; long chain (>=9 carbons); adamantyl; crown ethers; dipeptides; ethylene >=4 units; unbranched, unsubstituted carbon cycle (>=7 atoms); >=4 conjugated double (or aromatic) bonds; >=2 conjugated acetylenes; >=3 nitro groups; >=4 acetales, aminales, ureas, guanidines, carbamates etc.; aminales, hemiaminals (acyclic); quarternary ammonium salts; lactones; thiourea, thioamide, thioketones, and related, etc. (acyclic); more then two halo-substituents per aryl group; more then one nitro group per aryl group, one nitro and two halides per aryl.

(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3−26.

(18) We applied the following constraints: 250 < Mwt < 50; HBD < 5; HBA < 7; RotBonds < 8.

(19) Sertanty, Inc., San Diego, CA. Structure−activity data sets were selected from the Sertanty kinase knowledgebase[14] considering assay type, mode of ligand interaction, binding domain, kinase substrate, experimental assay conditions, and data consistency/reliability and used to develop several kinase eScreen QSAR models; http://www.sertanty.com/prod/software/escreens.html. Several enrichment case-studies are described in http://www.sertanty.com/ddd/SampleEnrichmentStudies.pdf.

(20) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(21) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(22) Deanda, F.; Stewart, E. L. Application of the PharmPrint Methodology to Two Protein Kinases. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1803−1809.

(23) Available Chemicals Directory. MDL Information Systems Inc., San Leandro, CA, USA; http://www.mdli.com/products/experiment/available_chem_dir/index.jsp.

(24) 28,000 structures were generated with an average Tanimoto similarity (using Daylight structural fingerprints)[30] of 0.5 to 0.6 with respect to the MDDR (65K), ACD (45K), a subset of the Sertanty kinase-active knowledgebase (23K) and commercially available (850 K) compounds from only a small subset of 5 reactions generating 14 virtual protocols; results unpublished.

(25) As an overview and general reference see the "Genetic Algorithms Warehouse": http://geneticalgorithms.ai-depot.com/Tutorial/Overview.html.

(26) Wessel, M. D.; Jurs, P. C.; Tolan, J.W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(27) Johnson, S. R.; Jurs, P. C. Prediction of Acute Mammalian Toxicity from Molecular Structure for a Diverse Set of Substituted Anilines Using Regression Analysis and Computational Neural Networks. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa, B.; Folkers, G., Eds.; Wiley-VCH: New York, 1997; pp 29−48.

(28) *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; Vol. 2.

(29) This model was developed using a training-set of 817 kinase-active compounds and approximately twice that number of diverse allegedly kinase-inactive compounds; at 0.5 cutoff in this binary model (trained with 0.0-inactive; 1.0-active) − 94.6% of actives and 94.7% of inactives are predicted correctly; this is a binary classifier, so no IC50 values were used in model development; results not published.

(30) 1024 bit structural fingerprints were calculated in Daycart 4.8, Daylight Chemical Information Systems; http://www.daylight.com/products/daycart.html.

(31) Dumas, J.; Sibley, R.; Riedl, B.; Monahan, M. K.; Lee, W.; Lowinger, T. B.; Redman, A. M.; Johnson, J. S.; Kingery-Wood, J.; Wilhelm, S. M.; Smith, R. A.; Bobko, M.; Schoenleber, R.; Ranges, G. E.; Housley, T. J.; Bhargava, A.; Scott, W. J.; Shrikhande, A. Discovery of a New Class of p38 Kinase Inhibitors. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2047−2050.

(32) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 MAP Kinase by Utilizing a Novel Allosteric Binding Site. *Nat. Struct. Biol.* **2002**, *9*, 268−272.

(33) Regan, J.; Breitfelder, S.; Cirillo, P.; Gilmore, T.; Graham, A. G.; Hickey, E.; Klaus, B.; Madwed, J.; Moriak, M.; Moss, N.; Pargellis, C.; Pav, S.; Proto, A.; Swinamer, A.; Tong, L.; Torcellini, C. Pyrazole Urea-Based Inhibitors of p38 MAP Kinase: From Lead Compound to Clinical Candidate. *J. Med. Chem.* **2002**, *45*, 2994−3008.

(34) Brown, D. S. Amide Derivatives, *PCT int. Appl.* WO0055153, 2000-Sep-21, AstraZeneca UK Limited.

(35) Selected reference: Bemis, J. E.; Buchanan, J. L.; Dipietro, L. V.; Elbaum, D.; Habgood, G. J.; KIM, J. L.; Marshall, T. L.; Geuns-Meyer, S. D.; Novak, P. M.; Nunes, J. J.; Patel, V. F.; Toledo-Sherman, L. M.; Zhu, X.; Armistead, D. M. Triazine Kinase Inhibitors. *PCT Int. Appl.* WO0125220A1, 2001-April-12, Kinetix Pharmaceuticals Inc.

CI0496853