JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

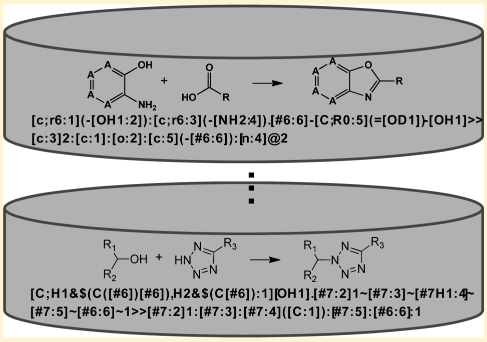
A Collection of Robust Organic Synthesis Reactions for *In Silico* Molecule Design

Markus Hartenfeller,[*,†] Martin Eberle,[†] Peter Meier,[†] Cristina Nieto-Oberhuber,[†] Karl-Heinz Altmann,[‡] Gisbert Schneider,[‡] Edgar Jacoby,[†] and Steffen Renner[†]

[†]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland
[‡]Swiss Federal Institute of Technology (ETH) Zurich, Switzerland

S *Supporting Information*

ABSTRACT: A focused collection of organic synthesis reactions for computer-based molecule construction is presented. It is inspired by real-world chemistry and has been compiled in close collaboration with medicinal chemists to achieve high practical relevance. Virtual molecules assembled from existing starting material connected by these reactions are supposed to have an enhanced chance to be amenable to real chemical synthesis. About 50% of the reactions in the dataset are ring-forming reactions, which fosters the assembly of novel ring systems and innovative chemotypes. A comparison with a recent survey of the reactions used in early drug discovery revealed considerable overlaps with the collection presented here. The dataset is available encoded as computer-readable Reaction SMARTS expressions from the Supporting Information presented for this paper.


[c;r6:1](-[OH1:2]):[c;r6:3](-[NH2:4]).[#6:6]-[C;R0:5](=[OD1])-[OH1]>>
[c:3]2:[c:1]:[o:2]:[c:5](-[#6:6]):[n:4]@2
[C;H1&$(C([#6])[#6]),H2&$(C[#6]):1][OH1].[#7:2]1~[#7:3]~[#7H1:4]~
[#7:5]~[#6:6]~1>>[#7:2]1:[#7:3]:[#7:4]([C:1]):[#7:5]:[#6:6]:1

INTRODUCTION

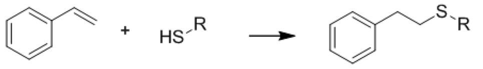
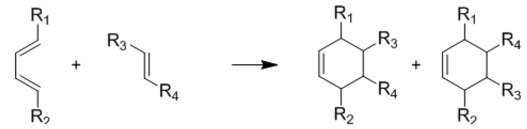
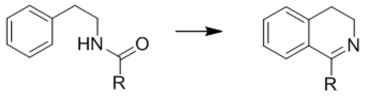
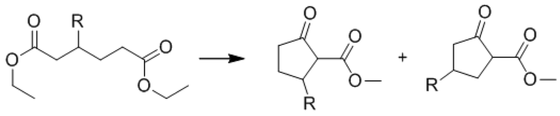
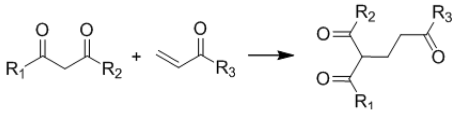
The generation of novel druglike compounds is a key factor driving the discovery of innovative and patentable medicines.[1] To accomplish this goal, a huge variety of chemical reactions finds application in the area of drug discovery. At the time of manuscript preparation, the Chemical Abstract Service (CAS) database[2] comprises ~34 million single-step and multistep reactions published in journal articles and patents, growing by 30 000−50 000 entries per week.[3] These numbers give an impression about the brisk activity in the field. However, the CAS database contains much redundancy, because the same reactions are applied to different reactants, and identical reaction steps are reported in different studies. The notion that there is a limited set of synthesis reactions accounting for a high fraction of the overall number of reactions carried out in drug discovery has been underpinned by two recent publications.[4,5] Both studies conducted statistical analyses on the usage of reactions in defined areas of the pharmaceutical industry. The review of Cooper et al.[4] focuses on the usage of reactions at the stage of lead optimization in the area of respiratory diseases at GlaxoSmithKline. Roughley and Jordan[5] took a snapshot of the reaction protocols reported in three different journals in 2008 by chemists from three pharmaceutical companies. Although slightly differing in the classification of reactions, both studies come to the comparable conclusion that "*...the medicinal chemist's perceived reliance on a small number of reactions (amide formation and Suzuki cross-couplings being the most often cited) is generally true (10 reaction types comprise almost two-thirds of all reactions)...*"[5] and "*...63% of the 4800 reactions fell into four reaction classes (alkylations, condensations (amides and sulfonamides), palladium-catalyzed couplings, and protecting group manipulations)...*".[4] Both author teams address these findings mainly to the pressure in industry to deliver compounds within a limited time frame, which calls for highly reliable and versatile reactions.[5]

Incorporation of synthesis expertise into computational methods for molecule design has the appealing advantage to direct the focus on regions of chemical space presumably enriched with synthetically feasible compounds. Chemical knowledge has been introduced to cheminformatics tools on different levels of abstraction.[6] Imitating bond rearrangements of known chemical reactions has a strong analogy to real-world chemistry. Applied on available molecular building blocks, this *in silico* approach to molecule construction can even deliver direct blueprints of hypothetical synthesis routes for designed compounds.

However, this presumes the availability of a collection of synthesis reactions encoded in a computer-interpretable format, which is rarely found in the literature. In 2003, Vinkers et al. published such a reaction dataset, together with their *de novo* drug design program SYNOPSIS.[7] The dataset consists of 75 reactions (three ring-closing reactions, predominantly carbonyl chemistry) encoded as Reaction SMARTS[8,9] reaction expressions. Two years later, Schuerer and co-workers published another reaction set, in combination with empirical reactivity and compatibility information.[10] The Supporting Information of this study contains 54 reactions (14 ring-closing reactions, 6 reactions for the removal of protecting groups) depicted as reaction schemes. They must be translated to a computer-interpretable format before being applicable to computational studies. For a previous study, we have assembled a reaction

Received: August 9, 2011
Published: November 11, 2011

Table 1. Rejected Reactions and Primary Reason for Rejection

	rejected reaction	primary reasons for rejection
1.		product prone to oxidation
2.		regio- and stereoselectivity
3.		limited variability of educts: good yields expected only for a few substitution patterns of phenyl moiety
4.		limited availability of educts; high reactivity of products
5.		low attractivity and high reactivity of products

set that represents the basis of the DOGS *de novo* design software.[11] This collection (58 unique reactions, 34 ring formations) served as the starting point for the reaction dataset we present here. The reactions are encoded in the line notation language Reaction-MQL.[12] Since it is not supported by publicly available software tools, this language requires a translation to a common format prior to use. To the best of our knowledge, except for the SYNOPSIS dataset, there is no collection of chemical reactions for computational molecule design publicly available that is encoded in a computer-processable format and can readily be used. While the SYNOPSIS reaction set has already proved to be capable of generating synthesizable bioactive compounds,[7] we see a possible deficiency in its potential to generate innovative chemotypes, because of the small number of ring-forming reactions.

The aim of this work is to assemble a focused set of reactions for *in silico* design of novel druglike compounds and make it publicly available in a computer-processable and ready-to-use format. Ideally, a reaction in the set would fulfill the following criteria: (i) be relevant for drug development, with respect to generated substructures; (ii) be robust; (iii) feature broad availability of starting material, and (iv) have a broad spectrum of tolerated functional groups. It is important to note that these rules represent soft criteria and will not necessarily all be fulfilled at the same time by each reaction. Selection of reactions has been conducted manually from the literature and in collaboration with medicinal chemists to foster its practical relevance for early drug discovery. We are well aware of the fact that the scope of the reactions (in terms of accepted reactants) is likely overestimated. Interference with neighboring groups by inductive and/or electronic effects is of importance in practice, and is only partially considered here. Construction schemes based on the reactions should be understood as a source of inspiration rather then readily applicable synthesis schemes.

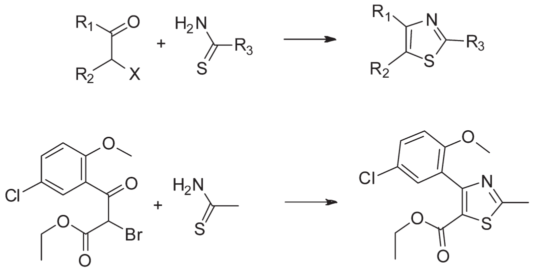
METHODOLOGY

Compiling of the Reaction Set. As a basis for collecting the reaction dataset, we started from a list of reactions that had been compiled for a previous project.[11] This previous collection was compiled in collaboration with chemists working in an academic environment. Although being intended for the same purpose of automated *de novo* design, it did not meet all of the requirements that we defined for the new reaction set. Three medicinal/combinatorial chemists at Novartis reviewed the collection individually, in order to shape the compilation toward robust reactions showing high potential for the design of combinatorial libraries. Table 1 displays examples of reactions which have been removed, together with the reason why they were disfavored for the intended purpose. Remaining reactions were partially modified by broadening or restricting the scope of accepted reactants. Reaction types deemed to be missing according to the intention of the data set and the defined soft criteria (*vide supra*) were added (e.g., the Sonogashira and the Grignard reaction). Finally, we scanned through the reaction sets provided by Vinkers et al.[7] and Schuerer and co-workers[10] to identify additional reaction types of interest for our purpose.

Encoding of Reactions. The reactions are encoded in the Daylight Reaction SMARTS[8,9] language. We decided to use this data format since it (i) allows for a detailed description of reaction centers, (ii) is widely supported by various software suites and programming toolkits for molecular modeling (both free and commercial), and (iii) is editable via a simple text editor. A reaction encoded as a Reaction SMARTS expression is a sequence of ASCII characters following a general pattern: a reactant side on the left is separated by the symbol "≫" from the product side on the right. In case there are more than one molecule on either side, they are separated by a period (".") (A.B ≫ C.D, where A and B are reactants and C and D are products). The reaction database described here only contains reactions of the form A ≫ B and A.B ≫ C, which is referred to as one- and two-component reactions, respectively. Only reaction partners that can structurally vary in some parts are explicitly considered as reactants (components) in the Reaction SMARTS expressions. Solvents, catalysts, and constant reagents are omitted from the Reaction SMARTS language. Reactants and products

Scheme 1. The Reaction SMARTS Expression (top) Only Describes Markush Representations of Reactants (middle, X = Cl, Br, I) and Leaves R-groups Variable. R-groups are Copied Unchanged to the Product Side upon Execution of the Reaction on Actual Reactants (bottom)

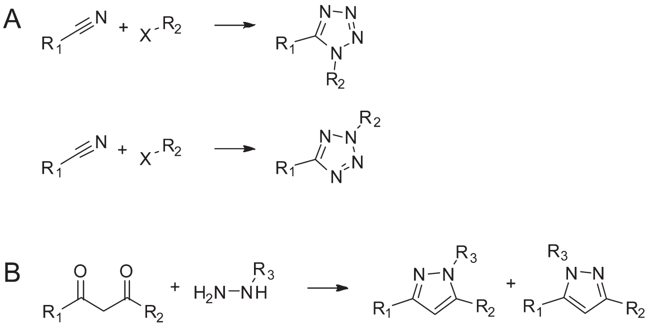
```
[#6:6]-[C;R0:1](=[OD1])-[CH1;R0:5](-[#6:7])-[*;#17,#35,#53].[NH2:2]-[C:3]=[SD1:4]>>
[c:1]2(-[#6:6]):[n:2]:[c:3]:[s:4][c:5]([#6:7]):2
```



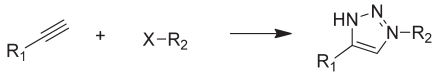
are defined manually and encoded as SMARTS substructure patterns.[8,9] Reactant substructures only cover atoms either directly involved in bond rearrangements (i.e., the *reaction center*) or deemed to be essential for the reactivity of the reaction center. An example from the dataset of an important functional group not involved in the reaction itself is the demand for a nitro group in the ortho or para position, relative to the halogen replaced in a nucleophilic substitution at a phenyl ring. The nitro group reduces the electron density at the carbon atom connected the halogen and promotes a nucleophilic attack at this position. The reaction expressions focus on parts of the molecules that are considered crucial for the reaction, while tolerating variation everywhere else. Upon executing a reaction on building blocks, those parts of a building block that are not explicitly described by the SMARTS expression (i.e., the R-groups of a Markush representation) are not modified and copied to the product side (see Scheme 1).

Stereocenters defined in the building blocks will be copied to the product side, as long as they are not part of the reaction center. Stereocenters generated by a reaction have no defined stereochemistry. The reason is that not all reactions can easily be stereo-controlled. In some cases, the control of stereochemistry is dependent on the molecular environment of the reaction center (e.g., the nature of the leaving group or steric hindrance by bulky side chains can shift ratios in one direction or the other). While it is generally possible to account for many of these effects in the Reaction SMARTS definitions, it would require different expressions for the same reaction. Although we are aware of the importance of stereochemistry for molecular recognition and, hence, bioactivity, we do not define the configuration of newly generated stereocenters in the presented reactions. In case it is desired for a certain project, it is straightforward to use the respective Reaction SMARTS expression as a template and explicitly define the configuration of the generated stereocenter on the product side. We prefer the strategy to leave generated stereocenters undefined and, in case a subsequent scoring step such as docking, three-dimensional (3D), shape or pharmacophore similarity will discriminate different stereoisomers, generate them in a post-processing step, e.g. by a conformer generator. It will then be up to an expert chemist to judge whether or not a certain enantiomer or diastereomer favored by the scoring scheme is feasible in the particular case. Another option for automated selection of preferred stereoisomers could be to perform a conformational analysis using a force field and select isomers based on calculated strain energies.

Scheme 2. (A) For Reactions Producing Regioisomers, the Formation of Each Regioisomer Is Described as a Separate Reaction Expression (X = Cl, Br, I; Step Implicitly Includes the Substitution of the Halogen with Azide Prior to Cyclization). (B) Reaction Expressions Featuring a Symmetrical Reactant Substructure Definition (Here, the 1,3-Dione) Inevitably Produce Regioisomers without the Need To Describe Each of Them Separately



Scheme 3. Example of a Reaction Including an Additional Implicit Reaction Step. This Triazole Synthesis Requires the Substitution of the Halogen (X) at R_2 with an Azide Group Prior to the Actual Ring Forming Step, Which Is Not Explicitly Done during Virtual Synthesis

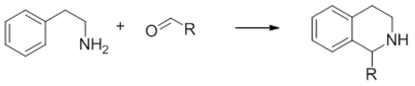
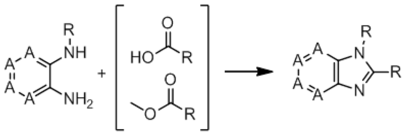
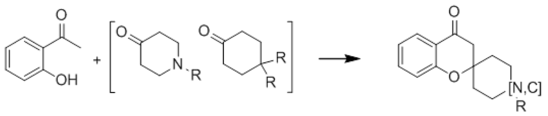
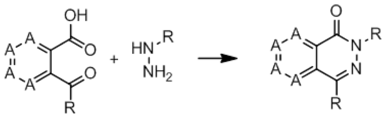
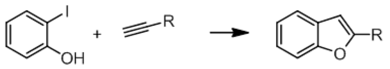


In contrast, reactions known to generate regioisomers have been split into distinct Reaction SMARTS expressions to account for the different regioisomers separately (see Scheme 2A). Regioisomers will most likely have a substantially different topological structure, which almost all computational molecule representations will discriminate. Furthermore, it is not straightforward to generate regioisomers in a post-processing step. Again, if a particular regioisomer is deemed to be of interest it is up to an expert to decide if it can be obtained selectively at the bench. In some cases the generation of different regioisomers does not even require an additional reaction expression. If one of the reactant substructure definitions is symmetrical it will match in different orientations by a substructure search, and the single reaction expression produces all regioisomer automatically (Scheme 2B).

Some reaction expressions implicitly include additional synthesis steps to make them directly applicable to a broader selection of reactants. An example of such a reaction is presented in Scheme 3. The halogen must be substituted by an azide group before the triazole ring can be formed. Building blocks featuring an azide group are rarely found in vendor libraries, because they tend to be instable over time. However, they can be produced from aliphatic halogens, which are commercially available in a large variety. By including this step implicitly, the reaction is applicable to a broader range of reactants without adding computational costs, and it still mirrors the practice at the bench.

Implementation. The code for the evaluation of Reaction SMARTS reaction expressions was implemented as a python script based on the RDKit computational chemistry toolkit.[13] The RDKit is an open source project and offers built-in functionality to interpret the Reaction SMARTS expressions.

Table 2. Examples of Ring Forming Reactions from the Reaction Collection[a]

	reaction scheme
1.	
2.	
3.	
4.	
5.	

[a] A = {C,N}; Ar = aryl.

■ DISCUSSION

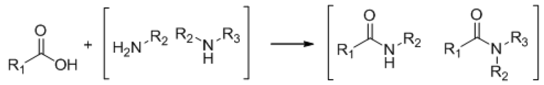
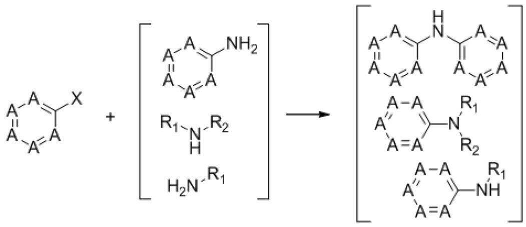
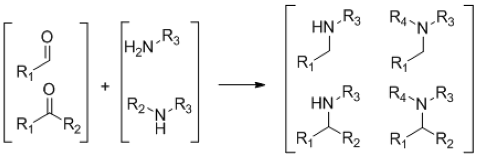
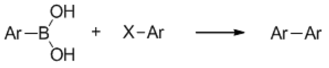
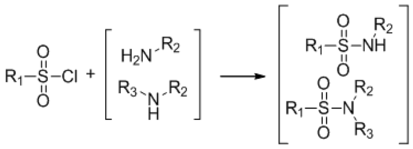
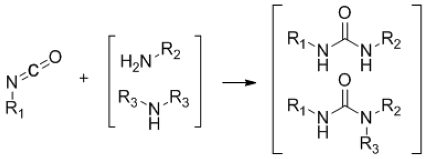
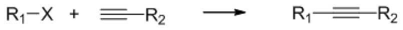
The final reaction dataset contains 58 generalized reaction schemes. A total of 29 of the schemes represent ring-forming reactions. The fact that half of the reactions are ring formations is intended to enhance the generation of novel chemotypes by constructing new ring systems. Table 2 displays a selection of ring-forming reactions included in the collection. The dataset splits into 56 two-component reactions and 2 one-component reactions. Multicomponent reactions are intrinsically more powerful in terms of product diversity, because they combine multiple reactants, and each of which can have a variable part.

A comparison between our dataset and the findings of the literature analysis by Roughley and Jordan[5] reveals that the 58 reactions already cover 3535 (48.3%) of the 7315 reactions steps under review (calculated by summing the number of reaction steps in reaction classes that can be mapped to reactions in our dataset). Please note that the literature survey did not serve as a primary source during the compilation process of the dataset; instead, it is used as a reference for retrospective comparison. Examples of prominent reaction classes covered by our collection (see Table 3) are *N*-acylation to amide (1165), *N*-arylation with Ar-X (458), reductive amination (386), Suzuki coupling (338), *N*-sulfonylation (163), *N*-acylation to urea (155), and Sonogashira (155). (Numbers in parentheses are taken from the survey by Roughley and Jordan[5] and represent the counts of reaction steps belonging to this class.) It is noticeable that 5 out of these 7 reactions include a primary or secondary amine as a reactant. A plethora of amines can be found in vendor catalogs of synthesis building blocks, which likely contributes to the fact that reactions based on these functional groups are frequently applied.

Reaction classes used frequently according to the survey but are not present in our set are *deprotections* and *protections*, accounting for 1319 and 225 reaction steps, respectively. Prediction of a protecting group strategy and explicit consideration of the underlying chemistry is outside of the scope of the dataset. Although being of high practical relevance, we think it is more efficient to elaborate this for a few selected candidate compounds of interest in a subsequent step. Integration of protection group chemistry into the *in silico* buildup process would likely slow it down, because of the considerable number of additional steps, which do not affect the final generated molecule. *N*-substitution with alkyl-X (390 reactions steps) represents another prominent reaction type which is not covered by our set. Although this type of reaction is prone to multiple substitution events, it might be worthwhile to consider a restricted version of the reaction for the future (e.g., restriction to secondary aliphatic amines as nucleophiles). We also do not account for reductions (406), oxidations (110), functional group interconversions (413), or functional group additions (78) in our reaction set. They were excluded because they are not within the scope of the dataset to assemble prototype compounds in a small number of steps. However, we are convinced that it can be worthwhile to consider some of these reactions in the future, e.g., either as a preprocessing step to expand the set of building blocks or as a post-processing routine after compound assembly in order to fine-tune functional groups. Ester condensation accounts for 46 reaction steps in the survey of Roughley and Jordan, which could be deemed a rather small number, given the vast availability of building blocks with suitable functional groups (carboxylic acid and hydroxy group). The reason probably is the instability of esters under physiological conditions,[14] which is why we decided to exclude the ester condensation (as well as the carbamate formation) from our dataset.

Although there is good rationale to exclude certain reaction types, the drawback of using a narrow range of synthesis reactions is obvious: less diversity in the underlying synthesis chemistry will likely result in less-diverse compounds being constructed.[4] Despite this valid concern, we think there are good practical

Table 3. Example Reactions from the Presented Dataset Corresponding to Reaction Classes Found to Be Frequently Used in the Survey by Roughley and Jordan[5a]

	reaction scheme	name
1.		N-acylation to amide
2.		N-arylation with Ar-X
3.		reductive amination
4.		Suzuki coupling
5.		N-sulfonylation
6.		N-acylation to urea
7.		Sonogashira

[a] Names given in the table have been taken from this study. Structures enclosed by square brackets are described by the same reactant SMARTS expression. A = {C,N}; Ar = aryl.

reasons that justify limiting a subset of reactions. First, the findings reported by Cooper et al.[4] and Roughley and Jordan[5] point to the fact that there seem to be reactions that are more relevant in the context of early drug discovery than others. The robustness of a reaction is a key feature making it broadly applicable for drug discovery efforts. Computational approaches based on such reliable reactions will likely have an improved chance to suggest synthesis routes that can then also be pursued at the bench. Second, even a small set of reactions —when combined with a building block collection of realistic size (10 000−50 000)—can span a combinatorial space that soon becomes too large to be fully enumerated with limited computational resources.[15] Focusing on a set of reactions relevant for drug discovery restricts the search space in a motivated way.

Instead of defining hard criteria on objective facts for the inclusion of reactions (e.g., yields, number of publications reporting application of a reaction, etc.), we decided to rely on the expert advice of experienced chemists involved in the design of chemical libraries for drug discovery. We expect this strategy to result in a dataset of high practical relevance and general acceptance by chemists. The authors are aware that a hand-compiled reaction set, such as the one presented here, is, by definition, subjective, and might change over the course of time and in the light of novel scientific findings. Rather than being a static set, the reaction database should be considered a snapshot of an ongoing process. We plan to update it based on experience gained by using it in practice.

CONCLUSIONS AND OUTLOOK

We present a collection of reliable reactions relevant for medicinal chemistry and encoded in computer-interpretable format. Only very few comparable sets of reactions for computational studies have been published so far. We are convinced that it is of benefit to incorporate synthesis knowledge into computational approaches such as small molecule *de novo* design in order to enhance the practical relevance of the results and achieve better acceptance by medicinal chemists. Another possible application could be the field of bioisosteric replacement, where the reactions can be used to select for building blocks not only mimicking steric and electrostatic properties, but also can be synthetically linked to the unchanged part of the molecule. The advantage over a simple substructure search for the respective functional group for linking is the ability to directly construct the virtual product. Changes introduced by replacing a part of the structure can thus be scored in the context of the complete molecule.

By publishing the reaction dataset, we hope to foster a discussion on the topic and receive feedback from the scientific community helping to improve the collection of reactions. A copy of the reaction dataset encoded in Daylight's Reaction SMARTS format can be downloaded as a text file from the Supporting Information. Schematic representations of the reactions and auxiliary information can be found in a separate PDF file in the Supporting Information; both are available free of charge via the Internet at http://pubs.acs.org or upon request from the authors.

Further enhancements of the dataset could include additional information about incompatibilities of reactions with certain functional groups, accounting for reactivity differences of functional groups, and an annotation of functional groups that might need to be protected if present in a reactant.

A detailed analysis of the reaction set's constructive potential and addressable chemical space is the subject of another publication currently in preparation, and will be published in the near future.

■ ASSOCIATED CONTENT

S Supporting Information. Reaction dataset encoded in Daylight's Reaction SMARTS format (Excel) and schematic representations of the reactions and auxiliary information (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: markus.hartenfeller@novartis.com.

■ ACKNOWLEDGMENT

We thank the NIBR Education Office for funding the NIBR postdoctoral fellowship of M.H. The authors are grateful to Thomas Zoller for his advice on compiling the reaction set. Gregory Landrum is acknowledged for his support on the RDKit.

■ REFERENCES

(1) Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H.-J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E. Recent trends and observations in the design of high-quality screening collections. *Fut. Med. Chem.* **2011**, 3 (6), 751–766.

(2) *Chemical Abstracts Service Database*; Chemical Abstracts Service: Columbus, OH, USA.

(3) Chemical Abstract Service. http://www.cas.org/expertise/cascontent/ataglance/index.html (accessed August 8th, 2011).

(4) Cooper, T. W. J.; Campbell, I. B.; Macdonald, S. J. F. Factors determining the selection of organic reactions by medicinal chemists and the use of these reactions in arrays (small focused libraries). *Angew. Chem., Int. Ed.* **2010**, *49*, 8082–8091.

(5) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54*, 3451–3479.

(6) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol. (N.Y.)* **2011**, *672*, 299–323.

(7) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F.; Heeres, J.; Koymans, L. M.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. SYNOPSIS: SYNthesize and Optimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765–2773.

(8) Daylight Chemical Information Systems, Inc., Laguna Niguel, CA, USA.

(9) *Daylight SMARTS Documentation*, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed August 8th, 2011).

(10) Schuerer, S. C.; Tyagi, P.; Muskal, S. M. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space. *J. Chem. Inf. Model.* **2005**, *45*, 239–248.

(11) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; *Schneider. DOGS: Reaction-driven de novo design of bioactive compounds*. Submitted.

(12) Reisen, F. H.; Schneider, G.; Proschak, E. Reaction-MQL: Line Notation for Functional Transformation. *J. Chem. Inf. Model.* **2009**, *49*, 6–12.

(13) RDKit Toolkit, www.rdkit.org (accessed August 8th, 2011).

(14) Satoh, T.; Hosokawa, M. The mammalian carboxylesterases: From molecules to functions. *Annu. Rev. Pharmacol. Toxicol.* **1998**, *38*, 257–288.

(15) Schneider, G.; Geppert, T.; Hartenfeller, M.; Reisen, F.; Klenner, A.; Reutlinger, M.; Hähnke, V.; Hiss, J. A.; Zettl, H.; Keppner, S.; Spänkuch, B.; Schneider, P. Reaction-driven *de novo* design: From virtual compound assembly via target profile prediction to chemical synthesis and biological testing of potential type II kinase inhibitors. *Fut. Med. Chem.* **2011**, 3 (4), 415–424.