JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Probing the Bioactivity-Relevant Chemical Space of Robust Reactions and Common Molecular Building Blocks

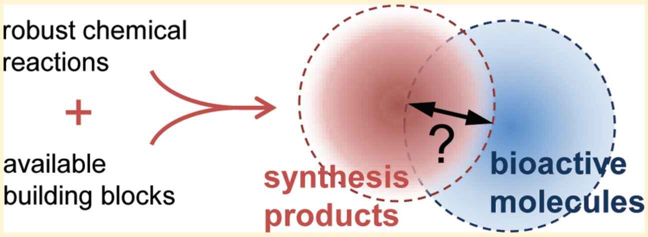
Markus Hartenfeller,*,[†] Martin Eberle,[†] Peter Meier,[†] Cristina Nieto-Oberhuber,[†] Karl-Heinz Altmann,[‡] Gisbert Schneider,[‡] Edgar Jacoby,[†] and Steffen Renner[†]

[†]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: In the search for new bioactive compounds, there is a trend toward increasingly complex compound libraries aiming to target the demanding targets of the future. In contrast, medicinal chemistry and traditional library design rely mainly on a small set of highly established and robust reactions. Here, we probe a set of 58 such reactions for their ability to sample the chemical space of known bioactive molecules, and the potential to create new scaffolds. Combined with ~26 000 common available building blocks, the reactions retrieve around 9% of a scaffold-diverse set of compounds active on human target proteins covering all major pharmaceutical target classes. Almost 80% of generated scaffolds from virtual one-step synthesis products are not present in a large set of known bioactive molecules for human targets, indicating potential for new discoveries. The results suggest that established synthesis resources are well suited to cover the known bioactivity-relevant chemical space and that there are plenty of unexplored regions accessible by these reactions, possibly providing valuable "low-hanging fruit" for hit discovery.



INTRODUCTION

The exploration of unexploited chemical space is a key goal of pharmaceutical research, both in industry and academia.[1] New compounds with innovative chemical structures represent opportunities to improve the pharmacological profile over existing compounds, and even address targets considered to be "undruggable" before.[2] In recent years, new paradigms to expand into the space of new, druglike, bioactive molecules have been devised and implemented,[3,4] e.g. diversity oriented synthesis[5] (DOS), biologically oriented synthesis[6,7] (BIOS), or macrocyclic molecules.[8,9] These strategies share the basic idea of generating molecules with increased structural complexity compared to currently available screening compounds (e.g., more stereocenters or higher fraction of sp^3-hybridized carbon atoms). It has been argued that this can increase the average performance in screening[10] (higher selectivity and protein binding rate) and development[11] (lower attrition rate). On the other hand, it has also been stated that the positive effect of raising the structural complexity of lead compounds likely has a peak at a certain complexity level and even drops beyond this point.[12] A drawback of most DOS and BIOS approaches is that they require substantial chemical synthesis development and, therefore, have not yet been taken up enthusiastically by pharmaceutical companies or commercial screening compound vendors. Moreover, the success of traditional libraries and also the more recent success stories of concepts based on chemically simple but very large libraries, as done e.g. by DNA encoded libraries,[13−15] proves that libraries made of chemically simple and tractable compounds might still be of high value for the discovery of bioactive compounds, despite the proven success of chemically more advanced approaches like DOS or BIOS. In this study, we explore how useful libraries based on established chemistry and commonly available building blocks can be for the discovery of chemical matter with the potential to be developed into future drugs.

We recently published a focused set of 58 organic chemistry reactions for in silico design of novel compounds.[16] By design, this collection has a strong bias toward synthesis methods commonly used in drug discovery. Here, we wanted to probe the data set's generative capabilities when combined with a library of ~26 000 readily available synthesis building blocks (see section Methods). The relevance of the chemical space spanned by the given synthesis resources (building blocks and reactions) for the discovery of new bioactive molecules is investigated by addressing two basic questions:

(1) Does the generated chemical space overlap with the relevant space of known pharmacologically active compounds?
(2) Is there potential for structural novelty in terms of innovative scaffolds within this chemical space?

RESULTS

Synthetic Accessibility Analysis. In order to address the first question (overlap with the chemical space of known

Received: December 21, 2011
Published: April 19, 2012

bioactives), we compiled three data sets of reference compounds with reported activity on human protein targets, the first containing 61 045 structurally diverse molecules from the GVK-BIO[17] database, a second collection of 871 traded drugs from the DrugBank database,[18] and a third data set comprising 115 compounds disrupting protein−protein interactions (PPI) from the 2P2I[19] and the TIMBAL[20] databases. For each data set, we determined the proportion of compounds which can be reconstructed based on 58 reactions and 26 043 available molecular building blocks (see section Methods for details on the data sets). For this purpose, a computational reconstruction routine was developed aiming to provide a possible synthesis route for a given reference molecule. Reconstruction of a reference molecule is done by explicitly building up the compound. First, the routine identifies building blocks which have the potential to be suitable for reconstructing the reference molecule. In order to achieve this, modified ("masked") versions of the building bocks are used as queries for a substructure search in the reference compound. *Masking* of a building block refers to the transformation of a functional group according to the chemical reaction in order to enable substructure searching. For example, the masking procedure referring to a Suzuki reaction substitutes a boronic acid moiety for an aromatic carbon atom. For each building block, all combinations of masked and unmasked functional groups are generated, and each of the resulting molecules is used as a query for a substructure search. The idea is to narrow down the number of building blocks which are used in the subsequent construction phase in order to reduce the search space (see section Methods for a detailed description). The number of reaction steps is not restricted.

Calculations were performed on a corporate computer cluster. In order to guarantee for a fair distribution of computational resources, the jobs were aborted after exceeding a maximum runtime (1 h). The results of the reconstruction analysis are summarized in Table 1.

Table 1. Accessibility of Compounds by in silico Reconstruction[a]

reference set	compounds	completed calculations	successful reconstructions	successful excluding 0-step syntheses
bioactives (GVK-BIO)	61045	38974	3675 (9.4%)	3599 (9.2%)
traded drugs	871	662	269 (40.6%)	96 (14.5%)
PPI-disrupting compd	115	87	15 (17.2%)	12 (13.8%)

[a] Numbers in parentheses denote percentage of completed calculations.

In silico reconstruction was successful for 3675 (9.4%) of the 38 974 GVK compounds that completed calculations. Traded dugs were reconstructed in 40.6% of the cases (269 out of 662 completed calculations). For PPI-perturbing molecules, 15 (17.2%) out of 87 calculated synthesis pathways were successful (data sets of all molecules that completed calculations are available from the Supporting Information).

The comparably high success rate for traded drugs can be explained by the high fraction of successful "zero-step" syntheses (173 out of 269), i.e. the reference molecule to construct is already available from the building block library. For comparison, the number of zero-step syntheses for the remaining two classes was only 76 out of 3675 (GVK) and 3 out of 15 (PPI). It is no surprise that the fraction of compounds already present in the set of commercially available building blocks is highest for marketed drugs, since they are of high interest as tool compounds and references for assay development. By excluding these compounds from the statistic, the success rate for traded drugs is reduced to 14.5%, which is in better agreement with the results found for the other data sets (Table 1).

In order to gain further insight into the scope of the chemical space accessible by the reaction data set, the results of the GVK-BIO data set were scrutinized.

Influence of Simple Molecular Properties on Reconstruction. The successfully reconstructed compounds from the GVK-BIO data set were compared with the unsuccessful examples in terms of simple molecular descriptors. The properties "number of atoms", "number of H-bond acceptors", "number of rings", and "number of aromatic rings" differed significantly between the two sets (p-values computed by Mann−Whitney U-test are $< 2.2 \times 10^{-16}$ for all properties). Visual inspection of the frequency distributions (cf. Supporting Information, part I) suggests the difference to be most distinct for the count of heavy atoms ($\text{mean}_{\text{successful}} = 27.6 \pm 6.6$, $\text{mean}_{\text{failed}} = 31.1 \pm 6.5$). Thus, successfully reconstructed molecules are smaller on average, which is in agreement with the assumption that larger reference molecules should usually be harder to construct. Compound sets also differ significantly (p-value $< 2.2 \times 10^{-16}$) in estimated synthetic accessibilities calculated by a method developed by Ertl and Schuffenhauer,[21] which rates compounds between 1 (easy to synthesize) and 10 (difficult to synthesize). The average score of successfully reconstructed molecules is 3.01 ± 0.53, compared to 3.55 ± 0.59 for compounds failing reconstruction. Successfully reconstructed ligands are estimated to be slightly easier to synthesize on average by this orthogonal method for assessing synthetic tractability. However, the difference between the distributions is too small to represent a strong hypothesis why some compounds failed to be reconstructed while others succeeded.

In contrast, the properties "number of H-bond donors" and Fsp^3 did not differ significantly at a significance level of 0.05 (p-values are 0.078 for the number of H-bond donors and 0.588 for Fsp^3). Fsp^3 describes a molecule's deviation from "flatness" as the ratio between sp^3-hybridized carbons and all carbon atoms and has been shown to correlate with physicochemical properties such as the melting point and aqueous solubility.[11] In addition, Fsp^3 describes an aspect of molecular complexity. A saturated ring system offers higher diversity of combinations of exit vectors than an equivalent unsaturated ring. Compounds failing reconstruction do not differ significantly in this aspect of molecular complexity from those which were successfully reconstructed.

An analysis of the compounds from the GVK-BIO data set which exceeded the maximum computation time indicates that these compounds are on average slightly larger (32.9 ± 6.6 heavy atoms) than the ones which completed the calculation (see property distributions, part I of the Supporting Information). The average synthetic accessibility score of compounds with aborted reconstruction runs is 3.84 ± 0.7. Although being higher than for the compounds which completed calculations, this finding cannot entirely explain the longer computational time. An inspection of the

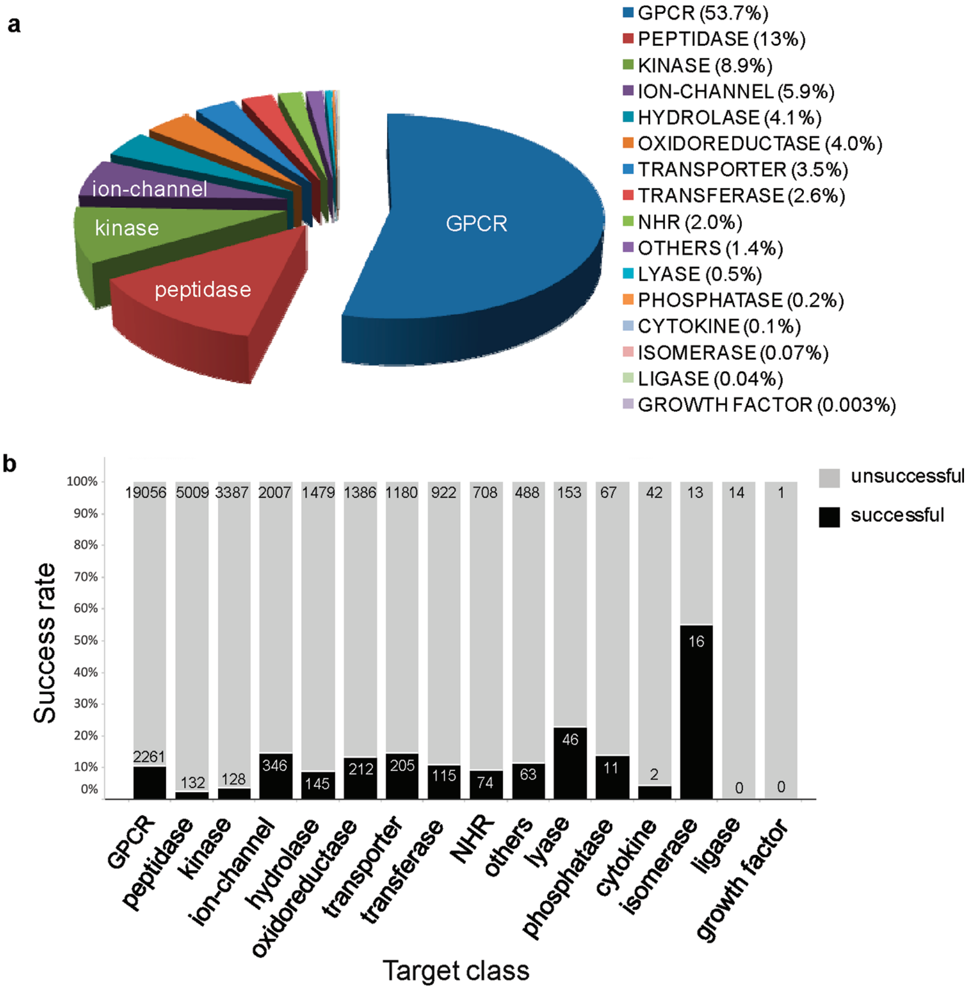


Figure 1. Target class coverage. (a) Distribution of ligands with respect to target classes in the GVK-BIO data set (compounds with completed reconstruction runs). (b) Black bars representing the percent success rates of compound reconstructions within each target family for the GVK-BIO data set. Numbers inside or above each bar represent successfully (black bars) and unsuccessfully (gray bars) reconstructed molecules. Only absolute identical reconstruction of the reference compound is counted as success. In cases where a ligand is active within more than one target family, it is counted multiple times. Target classes are sorted by number of ligands, decreasing from left to right (GPCR G protein-coupled receptor, NHR nuclear hormone receptor).

distributions suggests the difference between compounds with completed and aborted calculations to be most distinct for the Fsp^3 descriptor (0.32 ± 0.16 compared to 0.43 ± 0.15).

Coverage of Target Classes. Ligands of G-protein-coupled receptors (GPCRs) account for the largest fraction of the GVK-BIO compound set (53.7%), followed by peptidases (13%), kinases (8.9%), and ion-channels (5.9%) (Figure 1a). Success rates of compound reconstruction were calculated separately for each target class (Figure 1b).

Besides the two target classes represented by only very few ligands (ligases 14 ligands and growth factors 1 ligand), examples of successful reconstructions can be found for each protein family. For target families represented by low sample numbers (lyases, phosphatases, cytokines, isomerases, ligases, and growth factors), the extraction of a general trend is statistically not supported. However, there is a stable trend in the target families covered by more than 500 ligands. Except for peptidases (2.6%) and kinases (3.6%), success rates lie within a narrow range of 8.9–14.8%.

Synthesis Pathways. One intended application of the reaction data set is automated de novo design of bioactive compounds. Introducing chemical knowledge into de novo design should enhance the synthetic accessibility and practical relevance of suggested compounds. For this reason, we investigated the 3675 synthesis pathways leading to successful reconstructions of ligands from the GVK-BIO data set (each reconstructed ligand corresponds to exactly one synthesis pathway). In total, these synthesis pathways account for 8320 reaction steps. The frequency distribution of the number of synthesis steps is dominated by synthesis pathways with 1–3 reaction steps (87.2%, cf. the Supporting Information, part II). In 76 cases, the reference compound was already included in the building block library, i.e. a pseudo-synthesis of zero steps. Synthesis pathways of up to seven steps were suggested (two examples). It is obvious to assume a positive correlation between the size of a compound and the number of steps needed to construct it. In addition, it is also manifest that larger molecules have an increased probability to feature atoms which are not covered by one of the building blocks or bonds which cannot be formed by a reaction, and hence cannot be reconstructed. This might explain the dramatic drop of successful reconstructions with an increasing number of synthesis steps.

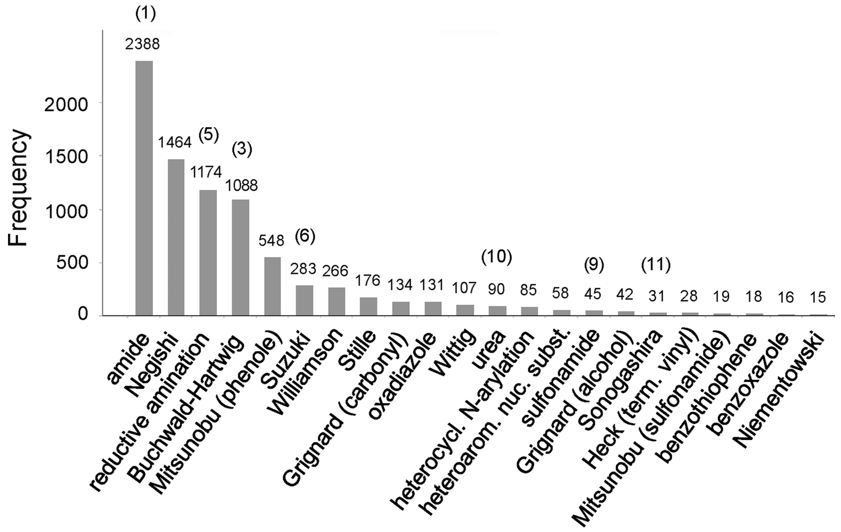
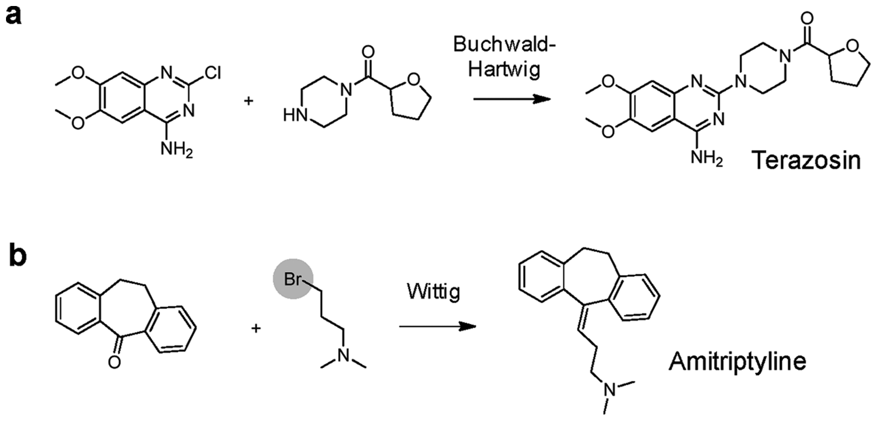


Figure 2. Frequency distribution of reactions. The numbers refer to the absolute occurrences of reactions in synthesis pathways of successful reconstructions (only including reactions used at least 15 times). Where available, reactions are named by their inventors; otherwise, they are named by mechanism or product (for details see ref 16). In the cases where a direct matching of categories was possible, additional numbers in brackets represent the rank of the respective reaction class in a survey on the frequencies of reactions in early drug discovery projects.[22]

Scheme 1. Suggested One-Step Synthesis Routes for Two Marketed Drugs[a]



[a](a) The synthesis of terazosin[24] has been proposed by the reconstruction algorithm exactly as reported in the literature.[25] (b) The *in silico* synthesis of amitriptyline[26] uses bromide (grey circle) instead of chloride[27] as a leaving group.

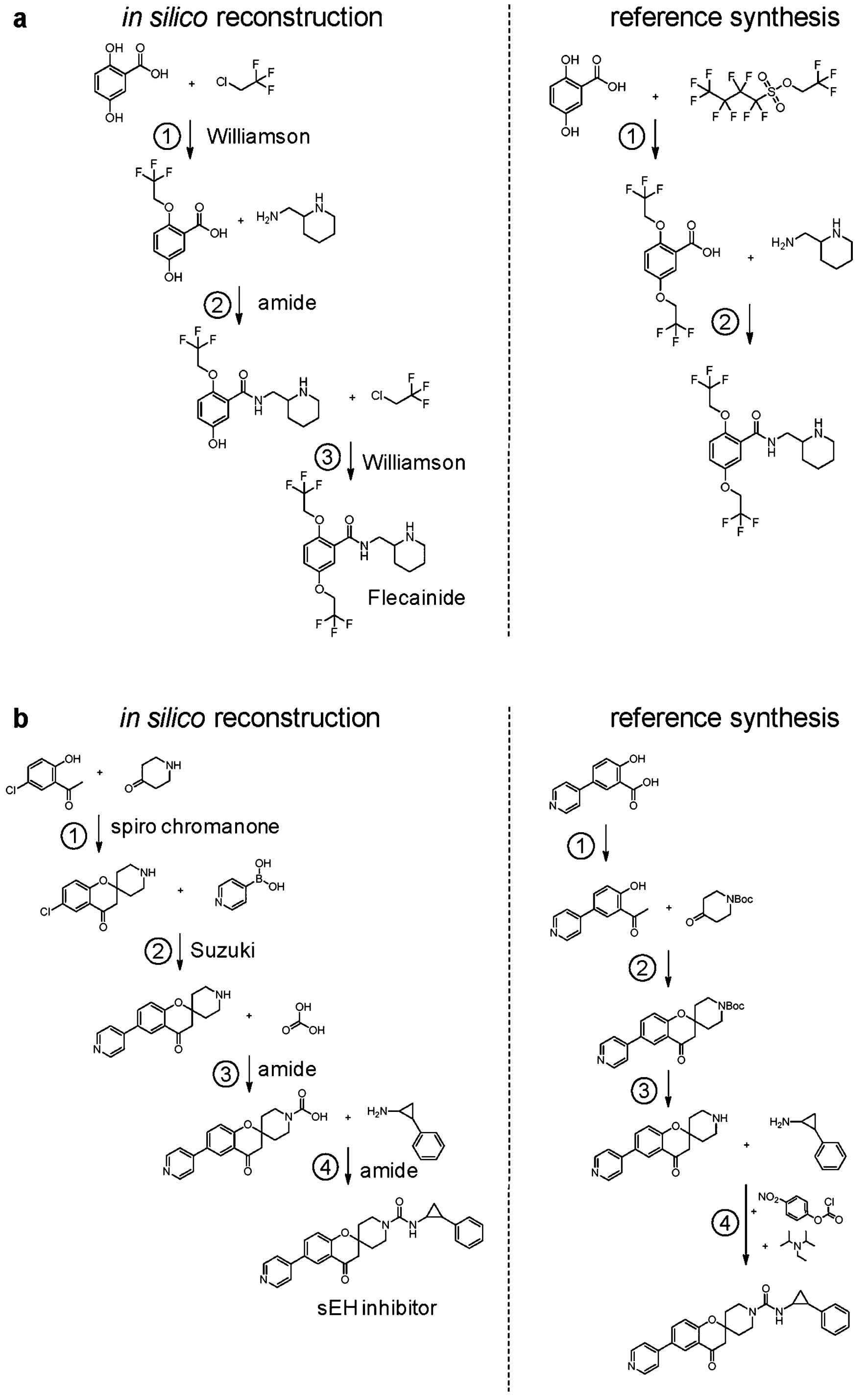
We also investigated the frequency of occurrence of each reaction over all successful reconstruction pathways from the GVK-BIO data set (Figure 2).

Amide bond formation between a carboxylic acid and a primary or secondary amine is by far the most frequently used reaction type (2388 reaction steps). This finding is in good alignment with the results of a survey on reaction usage in the early phase of drug discovery projects in the pharmaceutical industry by Roughley et al.,[22] which also ranks this reaction first (Figure 2). The frequent occurrences of reductive aminations, the Buchwald−Hartwig reaction and Suzuki couplings are also in agreement with the results of their study. The positive correlation between the popularity of a reaction and the availability of respective building blocks in the catalogs of commercial vendors[22] may presumably contribute to this result: reactions for which a broad range of building blocks is available are also more likely to be used in the reconstruction scenario. A comparison between our collection of reactions and the results of Roughley et al. and Cooper et al.[23] has been drawn before.[16] However, there are also contrary results. For example, the Negishi coupling is heavily used in the reconstruction pathways (1464 of 8320 reaction steps) but was found to be rarely applied in practice at the bench (category "other Pd-catalyzed reactions": 11 of 7315 reaction steps). This could either point at the fact that the definition of the reaction in the data set is not selective enough in terms of accepted reactants or that the Negishi reaction is considered unattractive for practical reasons.

Whether or not a synthesis pathway proposed by the reconstruction method is of actual practical relevance can either be shown by reproducing it in the lab, or by finding comparable synthesis routes reported in the literature. Since reconstructions were performed with compounds extracted from the literature, the latter approach is taken here. In the following, four exemplary reconstruction synthesis routes are compared to published synthesis strategies.

As a first example, the one-step synthesis of terazosin[24] has been reported[25] exactly as suggested by the reconstruction routine (Scheme 1a). In the case of amitriptyline,[26] the in silico synthesis pathway only slightly differs from the literature

Scheme 2. Comparison of Synthesis Routes Suggested by the Reconstruction Routine (Left) and Published Synthesis Routes (Right)[a]



[a](a) Synthesis[29] of the traded drug flecainide.[28] (b) Synthesis of an inhibitor of human soluble epoxide hydrolase (sEH). The presented synthesis pathway is extrapolated from a generic example given in the work of Shen et al.[30] including the inhibitor shown.

reference[27] by using a bromide instead of a chloride leaving group for a Wittig reaction (Scheme 1b).

The building block used in the reference synthesis (3-chloro-*N*,*N*-dimethylpropan-1-amine) is also available from the building block library and would be accepted as a reactant by the defined Wittig reaction. The selection of the bromo- instead of the chloro-substituted building block by the software is therefore arbitrary. We expect the reaction proposed by the software to be feasible as well.

The synthesis route of flecainide[28] represents a more complex example (Scheme 2a). The major difference between the synthesis pursued by the reconstruction routine and a synthesis pathway published in a patent[29] is the strategy of introducing two trifluoroethyl moieties. The software uses a Williamson ether formation which is split into two separate

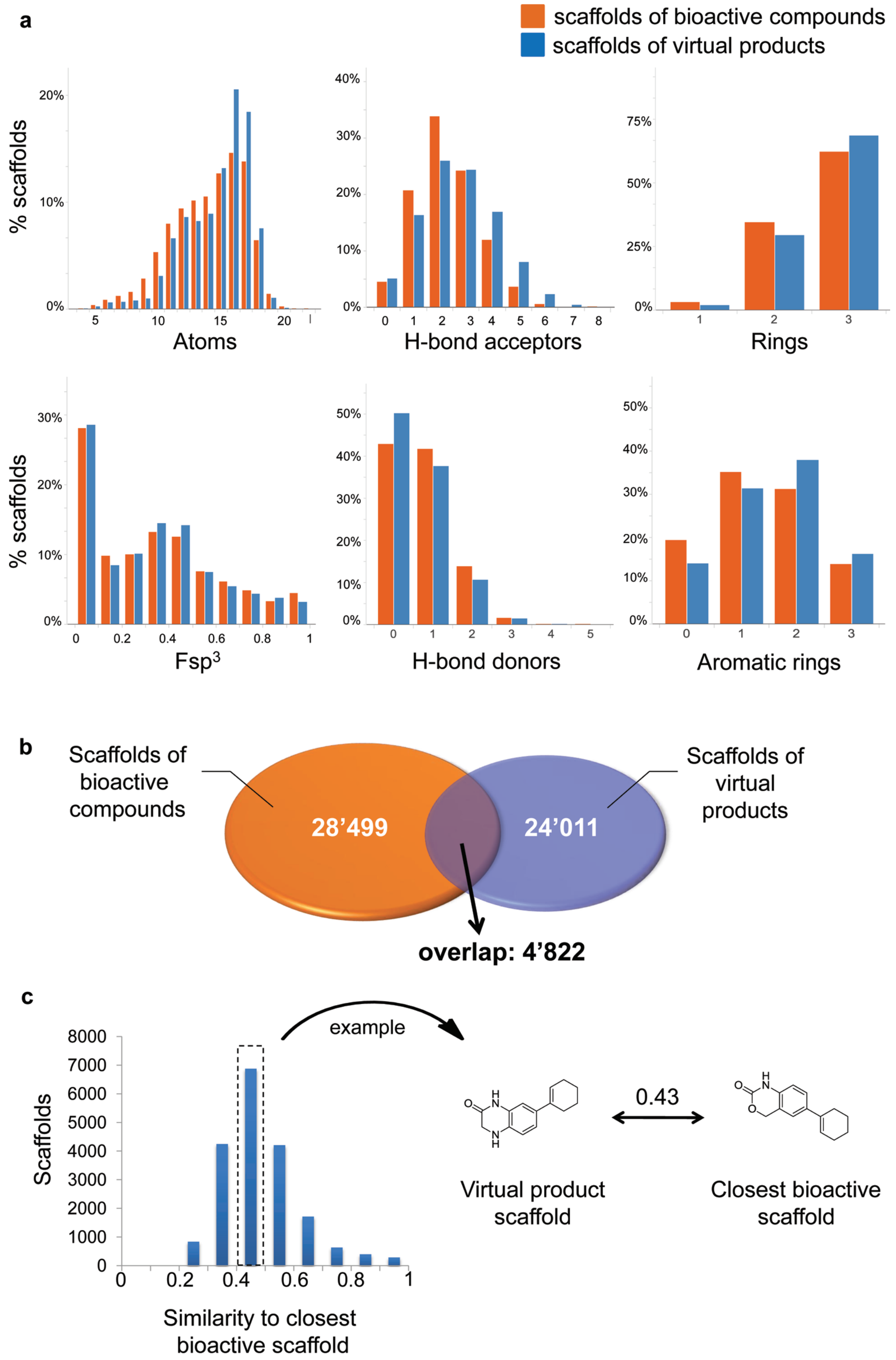


Figure 3. Comparison of scaffold sets. (a) Overlap of property distributions. (b) Overlap in terms of identical structures. (c) Distribution of structural similarities between scaffolds found exclusively in synthesis products and their most similar neighbor from bioactive scaffold set (left, Tanimoto index of ECFP_6 structural fingerprints). The highest populated bin (dashed box) is examplified by an exclusive synthesis product scaffold, together with its closest neighbor from the bioactive scaffold set and the computed similarity (right).

steps (steps 1 and 3) due to the implemented strategy. In practice, it would be preferable to combine the two steps into a single reaction step as it is shown in the reference synthesis. There, trifluoroethyl is introduced using 2,2,2-trifluoroethyl perfluoro-*n*-butanesulfonate. The remaining step (amide bond formation) is identical in both synthesis routes. Again, the suggested synthetic pathway shows considerable similarity to a validated synthesis.

As a last example, the synthesis of an inhibitor of human soluble epoxide hydrolase (sEH) reveals both the potentials and limitations of the synthesis pathways proposed by the in silico approach (Scheme 2b). Step 1 of the reference synthesis is the reduction of a carboxylic acid to the corresponding acetyl moiety. This step is not necessary in the reconstruction scheme due to the selected building block, which already features an acetyl group. The authors of the reference synthesis give no details on how the starting building block 2-hydroxy-5-(pyridin-4-yl)benzoic acid was obtained. Introduction of the pyridine side chain via a Suzuki coupling as proposed by the reconstruction algorithm (step 2) is a reasonable strategy. The ring closing reaction forming the spiro-piperidine moiety is identical in both syntheses (step 1 of reconstruction, step 2 of reference). A key difference between the two routes is the formation of the urea bridge between two aliphatic amines (steps 3 and 4 reconstruction, step 4 of reference). Formation of two amide bonds using carbonic acid as suggested by the reconstruction routine is not viable in practice. However, as exemplified in the reference synthesis,[30] synthesis strategies using two aliphatic amines form to form a urea bridge exist (step 4). This exemplifies how a proposed synthesis route can still serve as a valuable starting point for synthesis planning.

Scaffold Novelty. The second question we address is directed toward the potential of generating structural novelty based on the limited synthesis resources. In particular, we were interested in novelty on the level of molecular scaffolds. We extracted scaffolds from (i) 322 598 compounds from the GVK-BIO database reported to be active on human targets and (ii) 1 696 226 compounds originating from one-step syntheses based on the reaction data set and the building block library (see Methods section for details). Here, a scaffold is defined as a ring system of not more than three rings connected by not more than two nonring bonds (referred to as *compact scaffolds* in the following). The scaffold extraction protocol applied in this study is based on the scaffold networks method[31] and identifies all ring systems of a molecule matching our scaffold definition. Typically, this yields multiple scaffolds per molecule (see Methods section for details). Applying this routine resulted in two sets of 28 499 and 24 011 unique compact scaffolds from GVK compounds and from virtual one-step synthesis products, respectively. A comparison in terms of simple molecular properties revealed considerable similarity between the two scaffold sets (Figure 3a). The property distributions of scaffolds from the GVK compounds exhibit a small shift to smaller numbers for most of the descriptors (numbers of atoms, rings, aromatic rings, H-bond donors, and H-bond acceptors) in comparison to that of scaffolds from the one-step reaction products. This shift indicates a slightly smaller average size of the scaffolds retrieved from known bioactive molecules. However, it is evident that the regions occupied in property space by the two scaffold sets overlap considerably. Despite this finding, there are only 4822 identical scaffolds between the two sets (Figure 3b).

For each scaffold exclusively found in the virtual synthesis products, we computed the structural similarity (Tanimoto index, ECFP_6 fingerprints[32] in Pipeline Pilot[33]) to its closest neighbor from the bioactive scaffold set (Figure 3c). The average similarity evaluated to 0.49 ± 0.13 (average similarity of an exclusive bioactive scaffold to the closest synthesis product scaffold is 0.44 ± 0.13). The dominant part of calculated similarities lies between 0.3 and 0.6, which gives evidence for distinct structural differences (Figure 3c). Many of the suggested scaffolds are not present in the known bioactive compounds considered in this study. In addition, there is no close structural analog known for these "new" scaffolds.

DISCUSSION AND CONCLUSIONS

The aim of this work was to investigate the coverage of the bioactivity relevant chemical space and the potential for scaffold novelty of a limited set of established synthesis reactions and available building blocks. The limitation applied in this study might be deemed extreme, as there is a plethora of additional synthesis reactions and building blocks available which was not considered here. Nevertheless, we are convinced that the experiment covers a substantial part of the resources frequently used in drug discovery due to their robustness and wide applicability. Although efficiency in terms of time and costs may be the most dominant factor for focusing on a subset of available synthesis resources in practice, restrictions can also be imposed by other factors. For example, in DNA-encoded libraries,[34,35] (intermediate) products are labeled with nucleic acids, which requires reactions tolerating aqueous conditions.[13,36] Here, limits are imposed by the method. Against the background that genuine limitations exist in practice, we were interested in the chemical space that is still accessible by restricting synthesis resources to robust reactions and easily obtainable building blocks.

Despite restricted resources, a considerable fraction (9–15%) of the compounds for which calculations were completed is accessible via a hypothetical synthesis route for all three data sets investigated. This translates to a success rate of 6–11% for the whole data sets, assuming that all aborted calculations had failed to reconstruct the reference molecule. Successfully reconstructed ligands cover a broad range of pharmaceutically relevant target classes. The results for PPI-disrupting ligands—an emerging and challenging area for the development of low molecular weight modulators—hints at the data set's potential to reach into the chemical space of compounds being bioactive on emerging target classes. Reported success rates need to be considered in the context of the rigorous success criterion applied here, i.e. the exact reproduction of the reference compound. A shift of a single heteroatom of an aromatic ring by one position is sufficient to fail reconstruction. In addition, for technical reasons associated with this approach, there are examples where the algorithm fails to build a reference molecule that actually can be constructed with the given resources (see Methods section). Since this will only result in false negative examples, the reported success rates represent lower boundaries.

The frequencies of usage of reactions in successful reconstructions suggest that the reaction data set not only covers an essential part of frequently used reaction types,[16] but can also approximately mirror their relevance in practice. Detailed analyses of proposed synthesis pathways by comparison to literature examples show that the encoded reaction principles can translate into reasonable explicit synthetic steps. However, it must be noted that synthesis pathways followed by the reconstruction algorithm will not always represent direct blueprints for feasible synthesis routes. Nevertheless, they still might provide working hypothesis which can be transformed into viable synthesis plans by small modifications when reviewed by a medicinal chemist.

In conclusion, the set of reactions is capable of sampling the space of bioactive compounds over a broad range of target families, which answers the first question raised in the introduction.

In order to answer the second question, the potential for structural novelty in the accessible chemical space is evaluated

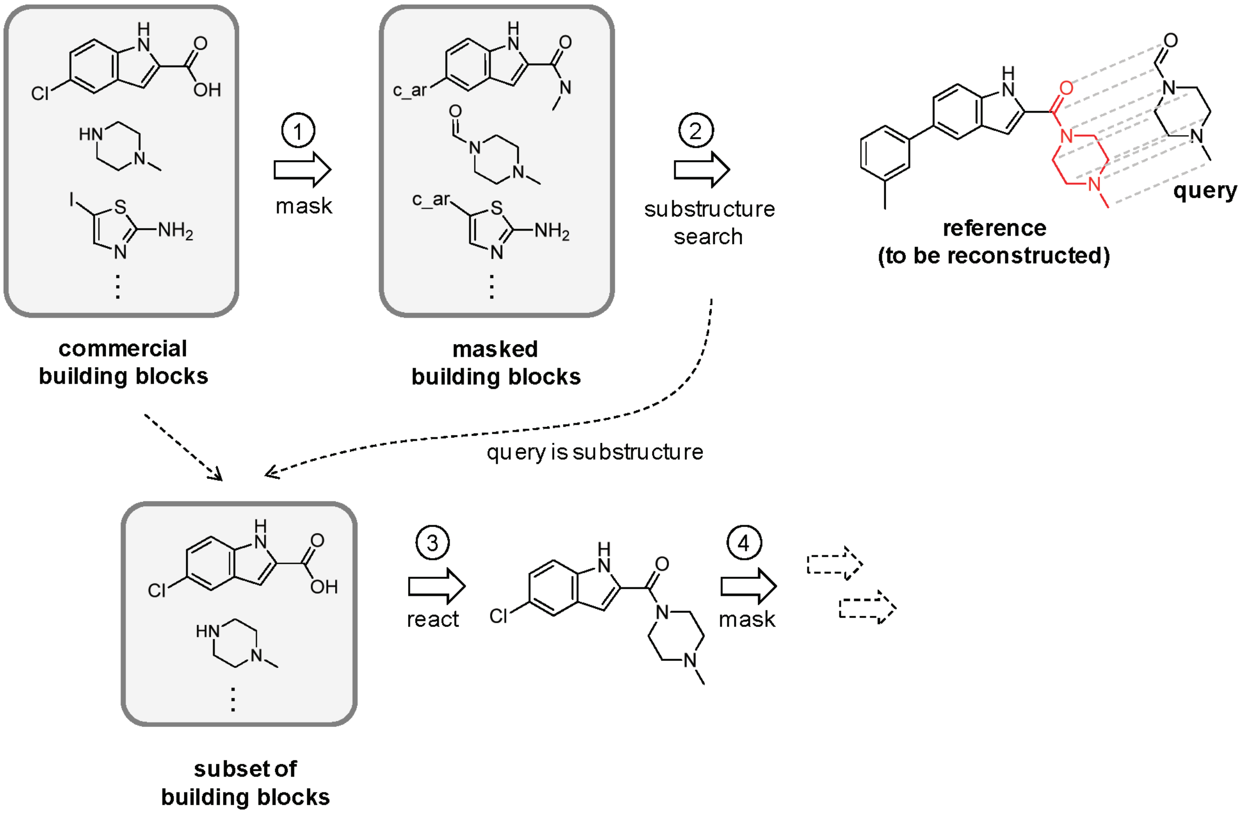


Figure 4. Reconstruction analysis. Numbers above the arrows denote the order of the steps. A successful substructure match in the reference compound during step 2 leads to the selection of the respective building block for step 3. Masking and coupling (steps 1 and 3) are based on the reaction data set.

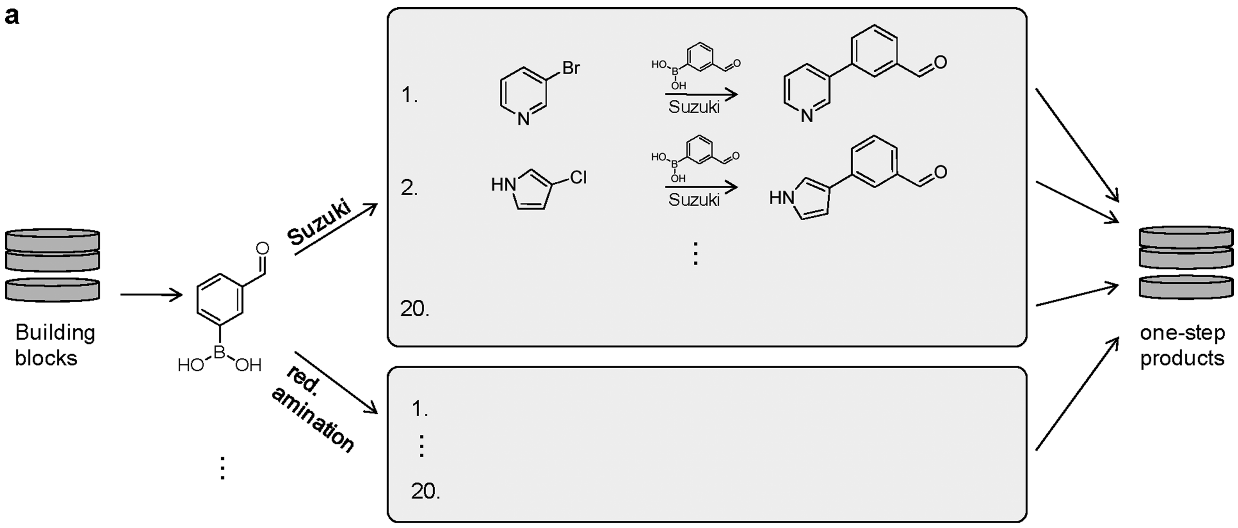
in terms of new molecular scaffolds. The definition of a "scaffold" applied in this study is focused on small, closely connected ring systems of not more than three rings linked by not more than two nonring bonds. Recently, it has been suggested that smaller ring systems (two rings in the example) are better-suited to evaluate the structural diversity of compound sets than conventional Murcko scaffolds, which were found to be too granular in representation.[37] For the same reason, we consider compact scaffolds a well suited representation for the assessment of chemical novelty. We argue that novelty found in a generalized representation is more meaningful because it offers fewer opportunities for differences than fine-grained representations (e.g., larger scaffolds or even complete compounds). In addition, such ring systems are of high relevance for the exploration of chemical space, as they represent common molecular starting points for the design of chemical series and the determination structure−activity relations.

It is notable that already a relatively small number of reactions, building blocks and generated reaction products generate a large number of compact scaffolds which cannot be found in a large set of known bioactive molecules. Almost 80% of the scaffolds from virtual synthesis products are not present in the ∼320 000 bioactive compounds considered. This suggests that the space of possible ring systems is not well covered by bioactive compounds explored so far—a conclusion which has been drawn from other studies as well.[38,39] We would like to stress that these results do not argue against the development of new synthesis strategies and application of more "exotic" reactions, which undoubtedly will be momentous for the development of innovative medicines. Instead, the results suggest that there is still plenty of unexplored chemical space accessible by established synthesis resources. These regions most probably contain valuable "low-hanging fruit", which one should keep exploiting in addition to the development of new, exploratory approaches. Future applications of the reaction data set in the field of computational molecule design can be a source for the discovery of promising compounds in these regions of chemical space.

■ METHODS

Data Sets. *Synthesis Reactions.* Organic synthesis reactions have been hand-compiled from the literature in collaboration with medicinal chemists. A detailed description of the data set has been published recently.[16] The data set comprises 58 reactions (29 ring forming reactions) and focuses on robust synthesis chemistry commonly applied in early phases of drug discovery projects.

Synthesis Building Blocks. The collection of synthesis building blocks was assembled from catalogs of ASDI[40] and Sigma Aldrich.[41] Only structures with a molecular mass between 30 and 300 Da were kept. All building blocks containing element types other than H, C, N, O, S, F, Cl, Br, I, and B were removed, as well as those having more than three aromatic rings or rings of more than eight atoms. The remaining set of 20 685 building blocks was extended by applying four functional group interconversion (FGI) reactions (aliphatic hydroxy → Br, aliphatic hydroxy → Cl, aromatic halogen → nitrile, isothiocyanate → thiourea). In case one of the functional groups was found in a building block, it was converted accordingly and the product was added to the data set. The original building blocks were also kept in the



b

```
for each of the 26,043 BBs:
        for each of the 58 reactions applicable to the current building block:
                select 20 compatible reaction partners from BB library by maximum structural diversity;
                execute current reaction with current BB and each of the 20 diverse BBs;
                add products to output library;
        end
end
```

Figure 5. Generation of one-step synthesis products. (a) Example of a building block (BB) with two applicable reactions. (b) Pseudo-code of the routine generating the one-step reaction products.

collection. This procedure expanded the collection of building blocks to 26 043 entries in total.

Bioactive Compounds. The GVK-BIO[17] database was filtered for compounds annotated with activity values (IC_{50}, EC_{50}, K_i, K_d) of <1 μM on at least one human target protein. This resulted in a selection of 322 598 compounds. For the accessibility analysis, this subset of GVK-BIO was subsequently filtered using additional criteria. Molecules were excluded if they (i) had a molecular mass >650 Da, (ii) contained any element other than H, C, N, O, S, F, Cl, Br, or I, (iii) contained a ring of more than eight atoms, or (iv) matched a set of in-house substructure definitions designed to exclude undesirable compounds from the in-house screening library. The remaining molecules were clustered according to their Murcko scaffolds[42] (including double bonds directly attached to the scaffold). From each of the resulting 61 045 clusters, the cluster center was added to the final compound set. Pairwise structural similarities were calculated as Tanimoto coefficients of Pipeline Pilot[33] ECFP_6 fingerprints.[32] This strategy was followed to reduce the number of compounds, while keeping the structural diversity in terms of scaffold variety. The final set comprised 61 045 unique compounds.

Traded Drugs. The "small molecule drugs" data set (6630 compounds) was downloaded from DrugBank[18] (http://www.drugbank.ca/downloads) and filtered for those tagged as "approved" (986 compounds). Duplicates, large molecules (>55 heavy atoms), entries exhibiting a ring of more than eight atoms, or any element type other than H, C, N, O, S, F, Cl, Br, and I were removed, which led to a set of 871 unique compounds.

PPI-Disrupting Compounds. Molecules perturbing protein–protein interfaces (PPI) from the 2P2I[19] database (http://2p2idb.cnrs-mrs.fr) and the TIMBAL[20] database (http://www-cryst.bioc.cam.ac.uk/databases/timbal) were combined into a single data set, and duplicate molecules were removed. The same structural filtering rules as described for the collection of traded drugs were applied, leading to 115 remaining molecules.

Reconstruction Routine. In order to decide whether or not a given reference compound is accessible by an in silico synthesis route based on the set of reactions and building blocks, we implemented a reconstruction strategy. Iteratively, the steps (i) masking, (ii) substructure searching, and (iii) coupling of fragments are repeated until a decision can be made. *Masking* introduces the minimal changes a certain reaction causes to a functional group (e.g., exchanges a boronic acid group with a single aromatic carbon atom in the case of a Suzuki coupling). A successful substructure search of a masked fragment (query) in the reference compound indicates that the unmasked fragment has to be considered as a potential building block for synthesis. Accordingly, the first step of an accessibility analysis is the identification of relevant building blocks via successful substructure matching of their masked equivalents in the reference compound (Figure 4, steps 1 and 2). In the subsequent step, the original building blocks are connected by applying the respective in silico reactions (Figure 4, step 3). By keeping track of the combination of masking reactions which led to a successful substructure match, the number of virtual intermediate products can be reduced to just the potentially relevant ones. A new iteration starts by masking the generated intermediate products (Figure 4, step 4), followed by

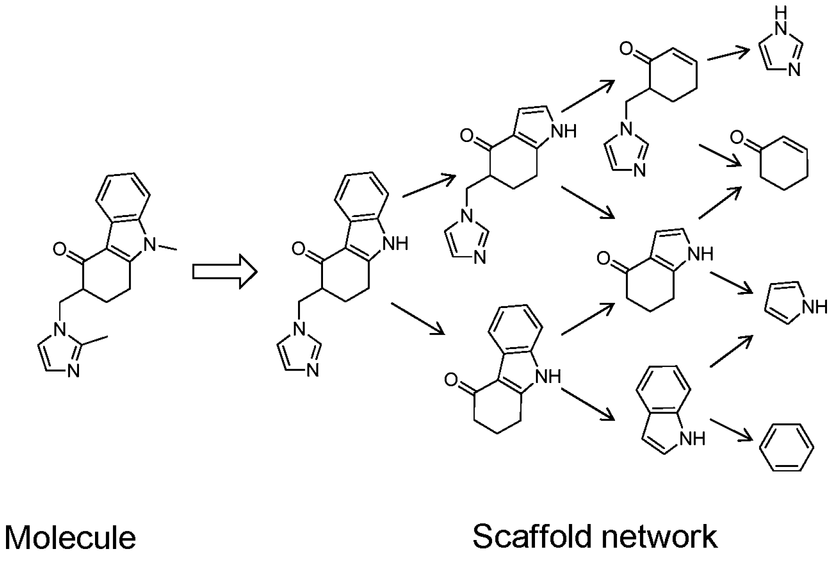


Figure 6. Scaffold network extraction. The scaffold network contains all ring systems (fused or linked) and single rings of a molecule. Atoms directly attached to a ring or linker chain by a double bond are kept, while all other exterior rings together with the connecting chains are iteratively removed. In the example, all generated ring systems of the scaffold network comply with our definition of a compact scaffold, except for the largest four-ring scaffold.

substructure matching in the reference. The construction process stops if none of the intermediates can be extended further or can be matched as a substructure of the reference (unsuccessful reconstruction), or if the reference compound has been reconstructed (successful). In case an atom of the reference can not be covered by at least one substructure match of any masked building block, the process terminates immediately with a negative result. The same is true if all building blocks matching the atom with the lowest building block coverage have been tried as synthesis starting points without success. If the synthesis of the reference compound is successful, the routine returns the corresponding synthesis pathway. Alternative pathways will not be explored once a successful synthesis route has been found.

There are cases where the reconstruction fails, although synthesis is possible based on the given resources. The masking approach will only cover changes made in one step. This means that in case a reaction step necessary for successful reconstruction assumes a certain reactant substructure that first has to be built by a previous reaction step, the masking concept fails. A second source of false negative results is the fact that some atom properties (e.g., aromaticity) depend on its chemical environment. In case the necessary environment (and consequently the property) of an atom of building block A will only be created by attaching building block B, A might not be considered a suitable building block. This happens if the molecular transformation of A by the masking approach is insufficient to create the chemical environment giving rise to the necessary atomic property.

Generation of One-Step Synthesis Products. In order to compare the scaffolds of known bioactive compounds with the scaffolds that can be generated based on the given building blocks and synthesis reactions, a pool of virtual one-step reaction products was generated in silico. Each building block undergoes each applicable reaction with a set of maximum 20 building blocks having the corresponding functional group in order to generate one-step reaction products. The 20 reaction partners are selected as a maximum diverse subset from all suitable building blocks in the library. For this purpose, all potential reaction partners are clustered by a hierarchical clustering routine based on pairwise Tanimoto coefficients of circular fingerprints (diameter = 4, equivalent to ECFP_4 fingerprints[32]) and a single linkage criterion, as implemented in RDKit, version 2010_12_1 (http://www.rdkit.org). The aim of the selection is to maximize structural diversity while keeping the overall number of products manageable. A maximum number of 20 reaction partners was found to be a reasonable trade-off. Figure 5 illustrates the strategy pursued and gives the pseudo-code for the implemented routine. Finally, all synthesis building blocks (26 043 entries) were added to the product library, since they can be regarded as zero-step products. This approach resulted in 1 696 226 unique compounds.

Scaffold Extraction. For the comparison between scaffolds found in bioactive compounds from the GVK-BIO database and the one-step reaction products, we extracted the scaffolds from each compound set separately. First, the compounds of a set were dissected by the scaffold network (SN) approach as described by Varin et al.[31] SN is an extension of the scaffold tree (ST) approach of Schuffenhauer et al.[43] and is calculated by a modified version of the Scaffold Tree Generator program of Scaffold Hunter.[44,45] Briefly, molecules are pruned by an iterative removal of peripheral rings and connecting linker chains. Fused rings (e.g., an indole ring system) are split, so that each ring of the fused system is detached separately. Each removal of a ring results in a distinct new molecule, which will be subject to the next pruning step (Figure 6). The SN of a molecule *m* will include all single rings, fused rings and linked ring systems that represent a substructure of *m*. For this analysis we focused on a subset of scaffolds. Only ring systems of one, two, or three rings connected by not more than two nonring bonds in total are considered (double bonds directly attached to a ring or a linker chain are not counted). This strict definition of a scaffold focuses on small, relatively rigid ring systems and excludes those with large, flexible linker chains. We refer to these ring systems as *compact scaffolds*.

Statistical Analysis. Statistical tests for significance of differences between distributions of molecular properties were performed using a two-sided Mann−Whitney U-test implemented in the software package R, version 2.8.1 (http://www.r-project.org/).

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Comparison of property distributions between molecules failing and passing reconstruction and the distribution of synthesis steps per reconstruction pathway (pdf) as well as the detailed results of the reconstruction analysis (three data sets, each as a separate text file). 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.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the Education Office of the Novartis Institutes for BioMedical Research for granting the postdoctoral fellowship of M.H. Wolfgang Zipfel is acknowledged for kindly providing access to additional computational resources for the reconstruction analysis. We thank John Peter Priestle for thorough proofreading of the manuscript.

■ REFERENCES

(1) Drewry, D. H.; Macarron, R. Enhancements of screening collections to address areas of unmet medical need: an industry perspective. *Curr. Opin. Chem. Biol.* **2010**, *14* (3), 289–298.

(2) Dandapani, S.; Marcaurelle, L. A. Accessing new chemical space for 'undruggable' targets. *Nat. Chem. Biol.* **2010**, *6* (12), 861–863.

(3) Jacoby, E.; Mozzarelli, A. Chemogenomic strategies to expand the bioactive chemical space. *Curr. Med. Chem.* **2009**, *16*, 4374–4381.

(4) 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. *Future Med. Chem.* **2011**, *3* (6), 751–766.

(5) Burke, M. D.; Schreiber, S. L. A planning strategy for diversity-oriented synthesis. *Angew. Chem., Int. Ed.* **2004**, *43*, 46–58.

(6) Kaiser, M.; Wetzel, S.; Kumar, K.; Waldmann, H. Biology-inspired synthesis of compound libraries. *Cell. Mol. Life Sci.* **2008**, *65*, 1186–1201.

(7) Wilk, W.; Zimmermann, T. J.; Kaiser, M.; Waldmann, H. Principles, implementation, and application of biology-oriented synthesis (BIOS). *Biol. Chem.* **2010**, *391* (5), 491–497.

(8) Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The exploration of macrocycles for drug discovery – an underexploited structural class. *Nat. Rev. Drug Discovery* **2008**, *7* (7), 608–624.

(9) Marsault, E.; Peterson, M. L. Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. *J. Med. Chem.* **2011**, *54* (7), 1961–2004.

(10) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (44), 18787–18792.

(11) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752–6756.

(12) Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856–864.

(13) Clark, M. A.; Acharya, R. A.; Arico-Muendel, C. C.; Belyanskaya, S. L.; Benjamin, D. R.; Carlson, N. R.; Centrella, P. A.; Chiu, C. H.; Creaser, S. P.; Cuozzo, J. W.; Davie, C. P.; Ding, Y.; Franklin, G. J.; Franzen, K. D.; Gefter, M. L.; Hale, S. P.; Hansen, N. J. V.; Israel, D. I.; Jiang, J.; Kavarana, M. J.; Kelley, M. S.; Kollmann, C. S.; Li, F.; Lind, K.; Mataruse, S.; Medeiros, P. F.; Messer, J. A.; Myers, P.; O'Keefe, H.; Oliff, M. C.; Rise, C. E.; Satz, A. L.; Skinner, S. R.; Svendsen, J. L.; Tang, L.; van Vloten, K.; Wagner, R. W.; Yao, G.; Zhao, B.; Morgan, B. A. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.* **2009**, *5* (9), 647–654.

(14) Buller, F.; Zhang, Y.; Scheuermann, J.; Schäfer, J.; Bühlmann, P.; Neri, D. Discovery of TNF inhibitors from a DNA-encoded chemical library based on diels-alder cycloaddition. *Chem. Biol.* **2009**, *16* (10), 1075–1086.

(15) Melkko, S.; Mannocci, L.; Dumelin, C. E.; Villa, A.; Sommavilla, R.; Zhang, Y.; Grütter, M. G.; Keller, N.; Jermutus, L.; Jackson, R. H.; Scheuermann, J.; Neri, D. Isolation of a small-molecule inhibitor of the antiapoptotic protein Bcl-xL from a DNA-encoded chemical library. *ChemMedChem* **2010**, *5* (4), 584–590.

(16) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093–3098.

(17) GVK Biosciences Private Limited, Plot No. 28 A, IDA Nacharam, Hyderabad 500076, India.

(18) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2010**, *39*, 1035–1041.

(19) Bourgeas, R.; Basse, M. J.; Morelli, X.; Roche, P. Atomic Analysis of Protein-Protein Interfaces with Known Inhibitors: The 2P2I Database. *PLoS ONE* **2010**, *5* (3), e9598.

(20) Higueruelo, A. P.; Schreyer, A.; Bickerton, G. R.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* **2009**, *74* (5), 457–467.

(21) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* [Online] **2009**, *1*, Article 8, http://www.jcheminf.com/content/1/1/8 (accessed March 5, 2012).

(22) 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.

(23) 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.

(24) Titmarsh, S.; Monk, J. P. Terazosin. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic efficacy in essential hypertension. *Drugs* **1987**, *33*, 461–477.

(25) da Silva, J. F.; Walters, M.; Al-Damluji, S.; Ganellin, C. R. Molecular features of the prazosin molecule required for activation of Transport-P. *Bioorg. Med. Chem.* **2008**, *16* (15), 7254–7263.

(26) Barbui, C.; Hotopf, M. Amitriptyline v. the rest: still the leading antidepressant after 40 years of randomised controlled trials. *Br. J. Psychiatry* **2001**, *178*, 129–144.

(27) Hudgens, D. P.; Taylor, C.; Batts, T. W.; Patel, M. K.; Brown, M. L. Discovery of diphenyl amine based sodium channel blockers, effective against hNav1.2. *Bioorg. Med. Chem.* **2006**, *14* (24), 8366–8378.

(28) Hudak, J. M.; Banitt, E. H.; Schmid, J. R. Discovery and development of flecainide. *Am. J. Cardiol.* **1984**, *53* (5), 17–20.

(29) Vigano, E.; Pizzatti, E.; Molteni, R.; Lanfranconi, S. *Process for the preparation of 2,5-bis-(2,2,2-trifluoroethoxy)-N-(2-piperidylmethyl)-benzamide (FLECAINIDE)*. U.S. Patent 6,599,922, filed August 8th, 2002.

(30) Shen, H. C.; Ding, F. X.; Wang, S.; Xu, S.; Chen, H. S.; Tong, X.; Tong, V.; Mitra, K.; Kumar, S.; Zhang, X.; Chen, Y.; Zhou, G.; Pai, L. Y.; Alonso-Galicia, M.; Chen, X.; Zhang, B.; Tata, J. R.; Berger, J. P.; Colletti, S. L. Discovery of spirocyclic secondary amine-derived tertiary ureas as highly potent, selective and bioavailable soluble epoxide

hydrolase inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19* (13), 3398–3404.

(31) Varin, T.; Schuffenhauer, A.; Ertl, P.; Renner, S. Mining for bioactive scaffolds with scaffold networks: improved compound set enrichment from primary screening data. *J. Chem. Inf. Model.* **2011**, *51* (7), 1528–1538.

(32) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(33) *Pipeline Pilot*, version 8.0; Accelrys, Inc.: San Diego, 2010.

(34) Buller, F.; Mannocci, L.; Scheuermann, J.; Neri, D. Drug discovery with DNA-encoded chemical libraries. *Bioconjug. Chem.* **2010**, *21* (9), 1571–1580.

(35) Kleiner, R. E.; Dumelin, C. E.; Liu, D. R. Small-molecule discovery from DNA-encoded chemical libraries. *Chem. Soc. Rev.* **2011**, *40*, 5707–5717.

(36) Scheuermann, J.; Neri, D. DNA-encoded chemical libraries: a tool for drug discovery and for chemical biology. *ChemBioChem* **2010**, *11* (7), 931–937.

(37) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* **2011**, *51* (9), 2174–2185.

(38) Ertl, P.; Jelfs, S.; Muehlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the rings: in silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573.

(39) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *J. Med. Chem.* **2009**, *52*, 2952–2963.

(40) ASDI Inc., Newark, DE.

(41) Sigma-Aldrich Co., St. Louis, MO.

(42) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(43) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree – Visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(44) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5* (8), 581–583.

(45) Renner, S.; van Otterlo, W. A.; Dominguez-Seoane, M.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided mapping and navigation of chemical space. *Nat. Chem. Biol.* **2009**, *5* (8), 585–592.