Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability

F. Chevillard and P. Kolb*

Department of Pharmaceutical Chemistry, Philipps-University Marburg, 35032 Marburg, Germany

Supporting Information

ABSTRACT: De novo drug design is widely assisted by computational approaches that enable the generation of a tremendous amount of new virtual molecules within a short time frame. While the novelty of the computationally generated compounds can easily be assessed, such approaches often neglect the synthetic feasibility of the molecules, thus creating a potential hurdle that can be a barrier to further investigation. Therefore, we have developed SCUBIDOO, a freely accessible database concept that currently holds 21 million virtual products originating from a small library of building blocks and a collection of robust organic reactions. This large data set was reduced to three representative and computationally tractable samples denoted as S, M, and L, containing 9994, 99 977, and 999 794 products, respectively. These small sets are useful as starting points for ligand identification and optimization projects. The generated products come with synthesis instructions and alerts of possible side reactions, and we show that they exhibit drug-like properties while still extending into unexplored quadrants of chemical space, thus suggesting novelty. We show multiple examples that demonstrate how SCUBIDOO can facilitate the search around initial hits. This database might be a useful idea generator for early ligand discovery projects since it allows a focus on those molecules that are likely to be synthetically feasible and can therefore be studied further. Together with its modular building block construction principle, this database is also suitable for structure−activity relationship studies or fragment-growing strategies.

INTRODUCTION

Chemical space is vast. The question is how to navigate it in order to identify ligands that can serve as modulators for pharmaceutically interesting targets. It seems likely that chemically not-yet-realized molecule sets hold many potent ligands for a variety of targets. At present, the in silico realm is the only place where we can hope to enumerate molecules that might be stable under ambient conditions. Such efforts have been undertaken, pioneered by Lederberg.[1] Currently, the most advanced development comes from the Reymond lab with their chemical universe database GDB,[2] which in its current incarnation enumerates virtual molecules containing up to 17 heavy atoms.

However, for all such virtual databases, the critical point is the actual synthesis of the generated molecules. Despite following strict chemistry rules, such molecules might turn out to be unsynthesizable with reasonable effort. This becomes a barrier in the initial stages of a lead-finding project, where a quick go/no-go decision is desired. In addition, automation of synthesis protocols is a currently intensively investigated topic of research,[3−5] and new open-innovation initiatives have arisen aiming at discovering novel chemical entities.[6] Thus, providing suggestions for further synthetic developments could improve such protocols even more.

Another problem of such enumerated databases is their sheer size: GDB currently holds 166 billion molecules,[7] which basically is computationally intractable except for ultrafast two-dimensional methods. For structure-based methods, such as docking, this is unfeasible at the moment.[8]

In order to advance on both topics, we have created the Screenable Chemical Universe Based on Intuitive Data OrganizatiOn (SCUBIDOO) and made it freely available to the general public. The current version was obtained by exhaustively reacting a set of building blocks with 58 highly reliable reactions. Such an approach is not completely novel[9−15] but has rarely been carried out entirely outside of an industrial framework.[16] The set of 58 reactions is the work of Hartenfeller et al.[17] and represents the most commonly used reactions in the pharmaceutical field. The authors compared their collection with a study by Roughley and Jordan[18] and showed that the 58 reactions cover 48.3% of the 7315 reaction steps described in this review. In a later publication, the authors investigated the coverage of chemical space afforded by their 58 robust reactions when applied to 26 043 common molecular building blocks.[12] They generated a limited number of one-step synthesis products by combining every building block with a maximum

Received: April 13, 2015
Published: August 17, 2015

DOI: 10.1021/acs.jcim.5b00203

of 20 reaction partners for each reaction. This protocol yielded a data set of 1 696 226 closed-source products and revealed that they were able to successfully reconstruct known ligands and sample the chemical space of bioactive compounds in a wide range of target families. Furthermore, they suggested that there is still a vast amount of unexplored bioactive space, which could contain "low-hanging fruit".

Going beyond this study, with SCUBIDOO we have now started to exhaustively react building blocks with each other in order to completely cover the chemical space thus accessible. This exhaustiveness comes at a price, however: even when starting with a relatively modest number of building blocks (~8000), this exhaustive reaction scheme yields a large number of lead-like molecules (more than 21 million) in the end. This is desired in the sense that the larger this number is, the greater is the amount of chemical space we can cover. However, many millions of virtual compounds make the utility for computational approaches dubious again. To address these divergent tendencies, we make use of stratified sampling to provide a representative subset of the database. Consequently, a user of our freely accessible database can obtain primary ligand candidates from a small and processable sample through virtual screening, pick those that can be synthesized with relative ease, and advance from hits efficiently by searching all analogues in SCUBIDOO. Since every molecule in the database comes with synthesis instructions, information about potential side reactions, and alternative synthetic pathways, it represents a fast and efficient way to start on a new target. SCUBIDOO can thus help to probe this unexplored potentially bioactive space more intuitively.

Moreover, we think that SCUBIDOO will help to fight "molecular obesity",[19] defined as the steady increase in the molecular size of drug candidates during medicinal chemistry development, since it facilitates starting a ligand discovery project focused on fragment considerations. Of course, the database as such can never be complete, but the concept is amenable to expansion.

In this article, we describe the development of the SCUBIDOO concept and a first database of 21 million compounds. We then show that the obtained database contains products comparable to currently existing drugs as well as databases of lead-like molecules of similar size. Still, SCUBIDOO extends into different quadrants of chemical space, as we demonstrate through principal component analysis (PCA). Moreover, we show the usefulness of SCUBIDOO through several examples in which we embark from close analogues of drug candidates or ligands and harvest even closer analogues or existing active compounds within a few mouse clicks. We also show that in those cases where synthetic information was publicly available, the reactions used to obtain these molecules match the ones suggested in SCUBIDOO.

■ METHODS

Reactions. The list of 58 reactions is provided in SMARTS notation in the study of Hartenfeller et al.[17] and in Table S1 in the Supporting Information.

Data Sets. *Reactants: Building Blocks.* Any product in SCUBIDOO is generated by combining a maximum of two building blocks. The initial set of 18 561 building blocks was downloaded from the ChemBridge Web site.[20] By means of a Python script written using the RDKit library,[21] routine filters were then applied to the building blocks to strip counterions and remove duplicates. In order to avoid overly complex reaction products that might necessitate more complex synthesis strategies, the reactant library was also pruned to narrow the range of generated products. The following criteria were used in the filters:

- *MW* ≤ 250 Da. Products will thus mostly be below 500 Da.
- Number of *rotatable bonds* ≤ 2. Since a reaction can introduce one or even two new rotatable bonds, this filter restrains the products to a low number of rotors (with a maximum of 6). Doing so makes the use of structure-based strategies such as docking more reliable, since the estimation of the binding energy of molecules with a high number of rotors is prone to fail.[22]
- Number of *chiral centers* ≤ 1. This filter was introduced with one goal: to facilitate synthesis. Since some of the 58 reactions introduce a chiral center into the resulting product, this limit on the number of chiral centers yields products with a maximum of three chiral centers.

Reactants: Analysis of Reagent Classes. In a recent study, Goldberg et al.[23] highlighted the importance of building block libraries in the interest of improving compound quality. They also introduced a classification of building blocks into 23 reagent classes based on functional groups. We used this class attribution in the analysis of the composition of the ChemBridge data set in this study, employing a Python script written using the RDKit library. Every building block was assigned to at least one reagent class but could belong to several classes. Using the SMARTS-encoded reagent class definitions as provided by the authors,[23] we reduced the list of 23 reagent classes to 22 by merging the *benzaldehyde* and *heterocyclic benzaldehyde* classes into a single *aromatic aldehyde* class. The list of 22 reagent classes is available in Table S2 in the Supporting Information. We note that this classification was done only to investigate the diversity of the library and had no influence on the reactions that each fragment was able to undergo.

Products: Screenable Chemical Universe. The filtered building blocks were then exhaustively reacted against each other via the 58 reactions using an in-house Python script written using the RDKit library. This procedure can be divided into three loops: the first one over all of the building blocks (B1), the second one over all of the reactions, and the third one over all of the building blocks (B2) (Figure 1). All of the products were charge-neutralized in order to simplify subsequent steps. Duplicate products were filtered according to their isomeric canonical SMILES notation and the reaction involved in the synthesis. We deliberately wanted to keep track

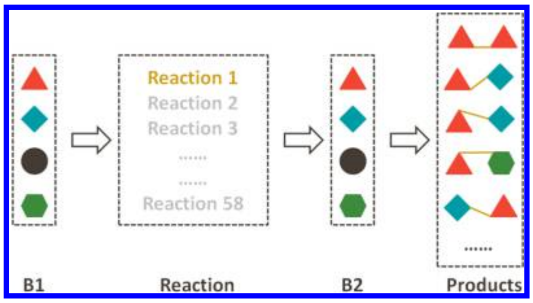


Figure 1. Schematic depiction of the creation of the SCUBIDOO database. B1 and B2 are building blocks which are then connected by all compatible reactions.

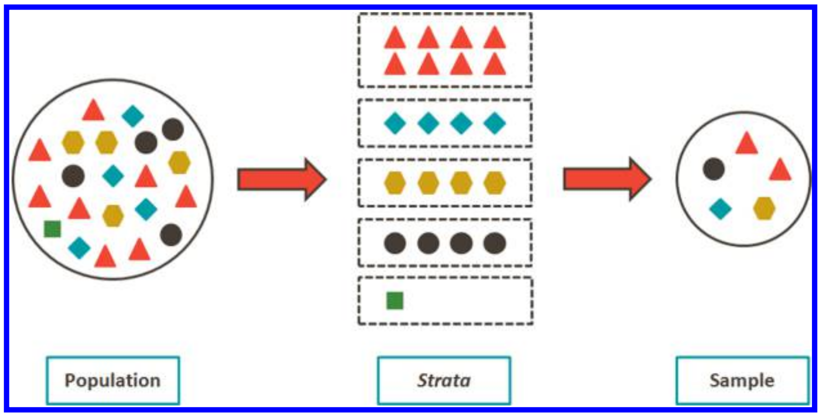


Figure 2. Stratified balanced sampling algorithm.

of multiple synthesis routes for the same product in order to display them on the subpage of each product as alternative synthetic routes. This is valuable information for chemists and increases the chances of synthesizing a particular compound. Afterward, the products were filtered using the PAINS filter level A[24] in order to get rid of compounds that have a high chance of generating artificial results during follow-up biological assays. Stereoisomers were generated using *flipper*[25] in order to enumerate all possible products in cases where a reaction introduced a new chiral center.

Representative Samples: Stratified Balanced Sampling. Since the present version and therefore also all future larger versions of SCUBIDOO are too big to be rapidly processed with structure-based approaches, we reduced it to three representative samples of different sizes (S, M, and L). This provides users with optimal sets for different applications. This procedure was done using the *cubestratified*[26] algorithm of the *balancedSampling* package[27] within the R statistics environment.[28] The stratified balanced sampling approach is a popular algorithm used for population surveys, allowing extraction of a representative sample. The algorithm consists of two stages (Figure 2). In the first stage, *stratification*, the studied set is divided into subgroups called *strata*. In this study, the strata are defined by the reactions, and each product is assigned to exactly one stratum. In a second step, *balanced sampling*[29] is applied within each stratum in order to select representative products. This selection is based on auxiliary variables defined as chemical descriptors (here, molecular weight, logP, number of H-bond donors, number of H-bond acceptors, and topological polar surface area were used) and aims to reflect the overall composition of each stratum. Furthermore, the sample size of each stratum is proportional to its total size. It is important to mention that balanced sampling does not guarantee that all of the strata defined initially are present in the final sample in cases where the strata sizes vastly differ. This algorithm is exceptionally fast even for huge amounts of data[30] and is thus well-suited for the processing of our screenable chemical universe (and its future larger incarnations).

DrugBank. DrugBank version 4.1 was downloaded from its Web site[31] for comparison purposes. Only approved and experimental drugs with molecular weights lower than 500 Da were kept. This led to a data set of 1510 molecules.

ZINC: Lead-like Subset. The lead-like data set was downloaded from ZINC.[32] Since this data set is quite large (more than 6 million compounds), a sample of 10 000 compounds was randomly selected using the *sample* function within the R statistics framework.

PDB: Ligands. All of the ligands present in the Protein Data Bank (PDB) were downloaded from its Web site.[33] The ligands were then filtered according to molecular weight (≤500 Da) and the number of rotatable bonds (≤6) in the interest of narrowing down the ligands to molecules close to the products of SCUBIDOO. This yielded 17 140 ligands.

Analogues of DB08235. The five analogues of DB08235 were prepared for docking using OMEGA,[34] with a maximum of 1000 conformers. The protonation states were defined using QUACPAC.[35]

Ligand-Based Application: Similarity Screening. A ligand-based screening strategy was applied in order to retrospectively assess the usefulness of SCUBIDOO. Two data sets were used in this comparison: DrugBank and all of the ligands extracted from the PDB. Each of these data sets was compared to the three samples of SCUBIDOO using FCFP4 fingerprints.[36] For each product in each sample, the most similar drug or ligand according to the Tanimoto coefficient was retrieved. All pairs with a Tanimoto score higher than 0.6 were visually inspected in order to identify the representative examples described in this article.

Synthetic Accessibility. To assess the synthetic feasibility of the products within SCUBIDOO with an alternative method, the synthetic accessibility (SA) score[37] was computed for each of the products using an RDkit-based Python script. SA score estimation is based on fragment contributions and a complexity penalty (chiral centers, weight, large rings). SA scores range between 1 and 10, with 1 indicating a simple molecule that should be easy to make and 10 representing a complex molecule that is likely to be hard to synthesize.

Principal Component Analysis (PCA). DrugBank, the lead-like subset of ZINC, and the SCUBIDOO S sample were compared using the PCA function as implemented in the R statistics environment. The descriptors used were molecular weight, logP, number of H-bond donors, number of H-bond acceptors, number of rotatable bonds, topological polar surface area, and the Bertz index,[38] which estimates the molecular complexity.

Web Interface. The database is freely accessible at www.kolblab.org/scubidoo. All of the chemical descriptors shown for each molecular entity are computed using the RDKit library. The partition coefficient, logP, is predicted using Crippen's approach.[39] TPSA represents the topological polar surface area of the molecule.[40] Two different searches are available: by product or by building block. These are described in more detail below:

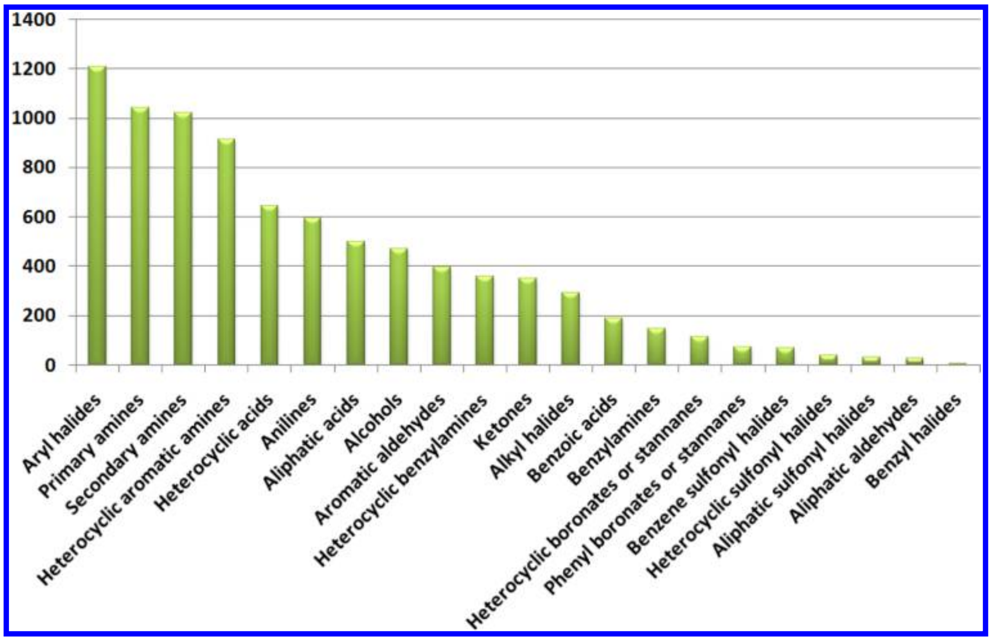


Figure 3. Frequencies of reagent classes in the ChemBridge building blocks data set.

Product Search. One can search for a product in SCUBIDOO using either its ID or its SMILES string. A comprehensive array of information is displayed when the product is retrieved: molecular descriptors, synthetic route, building blocks involved in the formation of the product, possible side reactions, and alternative synthetic routes. An alternative synthetic route is defined as a possibility to obtain a given product using a different reaction or a different pair of building blocks. A simple color classification has been implemented in order to quickly make the user aware of potential problems. *Red* products are ones that still contain reactive features (i.e., an electrophile and a nucleophile) and thus have the potential to react further. *Orange* products are ones for which the building blocks used in the synthesis are involved in more than one reaction (i.e., side reactions might occur). *Green* products are compounds that do not fall into the two aforementioned categories. Reactive features are retrieved using a Python script written using the RDKit library. Nucleophiles are defined as amines, alcohols, and thiols, while electrophiles were defined as acids, halides, and carbonyls. All of the functional groups are encoded as SMARTS and are provided in Table S3 in the Supporting Information.

Building Block Search. A building block search also displays a multitude of information: molecular descriptors, the top four analogue building blocks along with their Tanimoto scores using MACCS fingerprints, and all of the products in SCUBIDOO based on this building block grouped by reaction. For the latter option, the user has the possibility to download the selected products in SMILES format.

■ RESULTS

Creation of the Data Sets. *Filtering of the Building Block Library.* The initial library contained 18 354 building blocks. In a first filtering stage, all of the counterions were stripped and duplicates were removed, yielding 14 831 building blocks. Then only building blocks with molecular weights lower than 250 Da were kept, leaving 13 678 entities. Removing building blocks with three or more rotatable bonds reduced the library to 8006 entities. Applying a last filter allowing zero or one chiral centers resulted in a final building block library of 7805 molecules.

Creation of the Product Database. The 7805 building blocks were reacted against each other, generating 17 538 385 products. The computational part was carried out on a cluster of 192 CPUs and took less than 12 h. Duplicate products were removed, yielding 14 215 760 products. Then the PAINS filter level A was applied, leading to a reduction to 14 072 131 products. Afterward, stereoisomer generation was carried out, giving rise to a final database of 21 035 460 products.

Creation of Representative Samples. The 21 million products were reduced to three representative samples of different sizes (S, M, and L) using the stratified balanced sampling algorithm. All of the products were regrouped into strata by reactions, leading to 45 strata, i.e., 13 reactions never occurred. The never-occurring reactions are based on building blocks that are not present in the currently used library. The list of reactions that are not present in SCUBIDOO is provided in Table S4 in the Supporting Information. Next, during balanced sampling, the representative products for each stratum were selected using chemical descriptors (molecular weight, logP, number of H-bond donors, number of H-bond acceptors, and topological polar surface area) as auxiliary variables. In the end, the S, M, and L samples contained 9994, 99 977 and 999 794 compounds, respectively. For analysis, only the S sample was used. Only the L sample contained at least one representative of each reaction; the S sample was missing four reactions and the M sample two reactions.

Analysis of the Data Sets. *Analysis of the ChemBridge Building Block Library.* The breakdown of the ChemBridge building block library by reagent class is displayed in Figure 3. The most frequent reagent classes, which are primary, secondary, and heterocyclic aromatic amines, heterocyclic acids, and aryl halides, are consistent with the most frequent reactions used during the creation of SCUBIDOO. Indeed, the Schotten−Baumann amide, Buchwald−Hartwig, reductive amination, and Negishi reactions, which are the top four reactions employed in SCUBIDOO, require the aforementioned reagents.

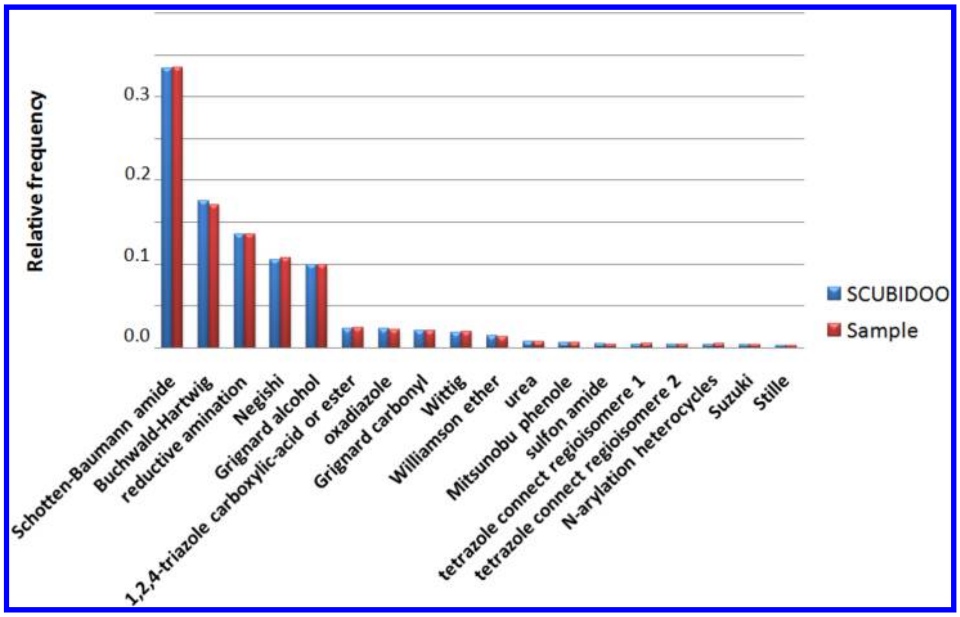


Figure 4. Relative frequencies of the reactions used in SCUBIDOO (blue) and the S sample (red). Only reactions employed at least 50 000 times are represented here.

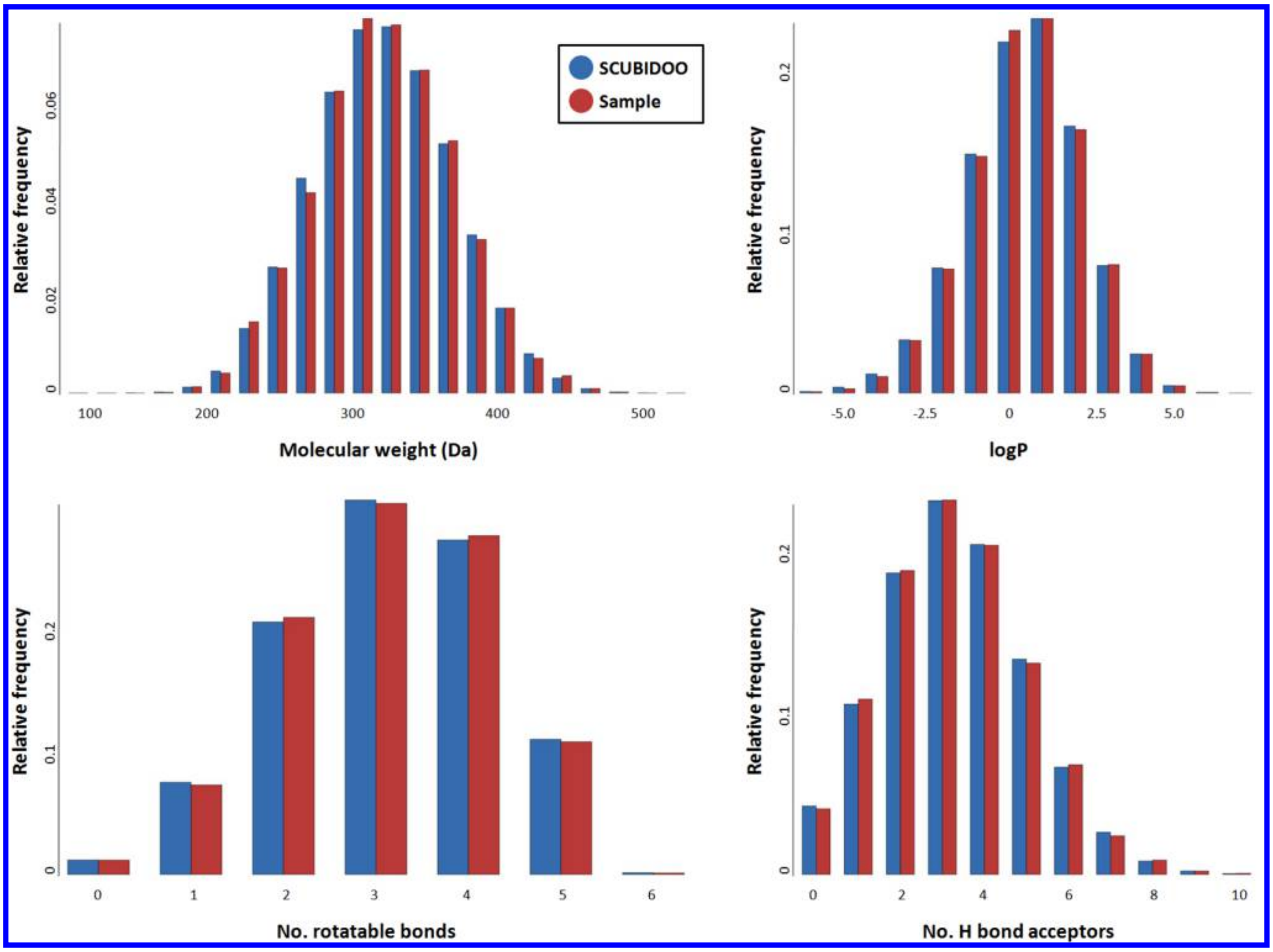


Figure 5. Relative frequencies of the descriptors molecular weight (upper left), logP (upper right), number of rotatable bonds (lower left), and number of H-bond acceptors (lower right) in SCUBIDOO (blue) and the S sample (red).

Frequencies of the Reactions. During the first stage of stratified balanced sampling, all of the products from SCUBIDOO were regrouped by reaction in order to define the strata. To ensure that the frequencies of the reactions

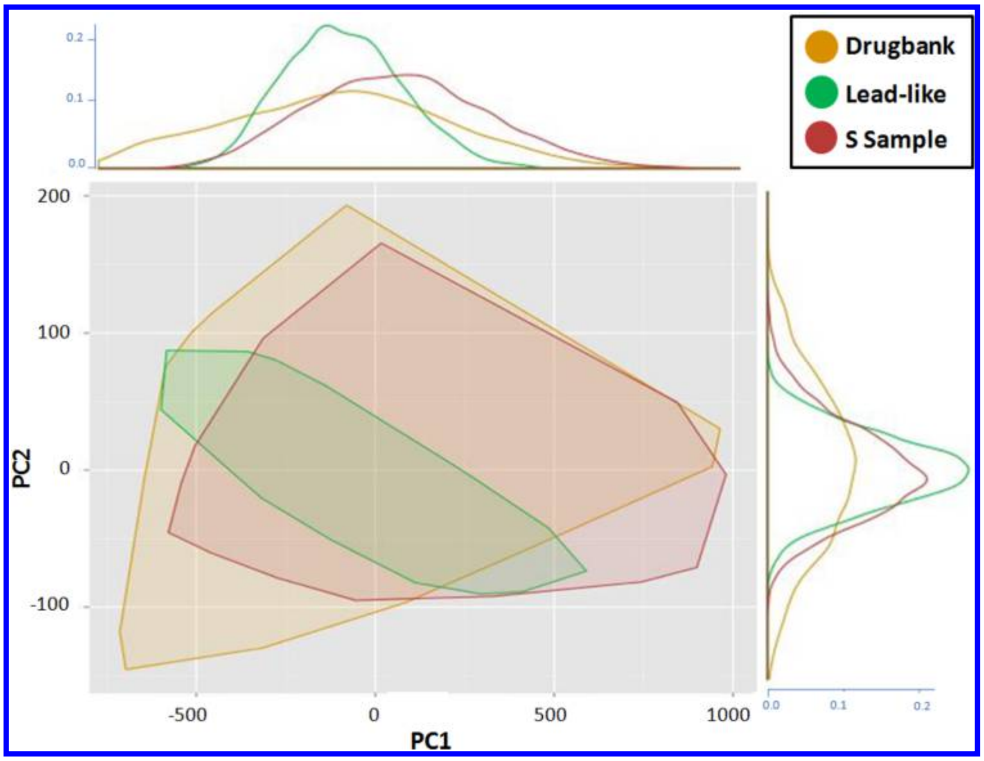


Figure 6. Principal component analysis of DrugBank (orange-yellow), the lead-like subset (green), and the S sample (red). The first principal component explains 34.9% of the total variance, and the second explains 28.3%. The two principal components thus cover 63.2% of the total variance. The marginal histogram on each axis represents the density distribution of each data set according the same color code.

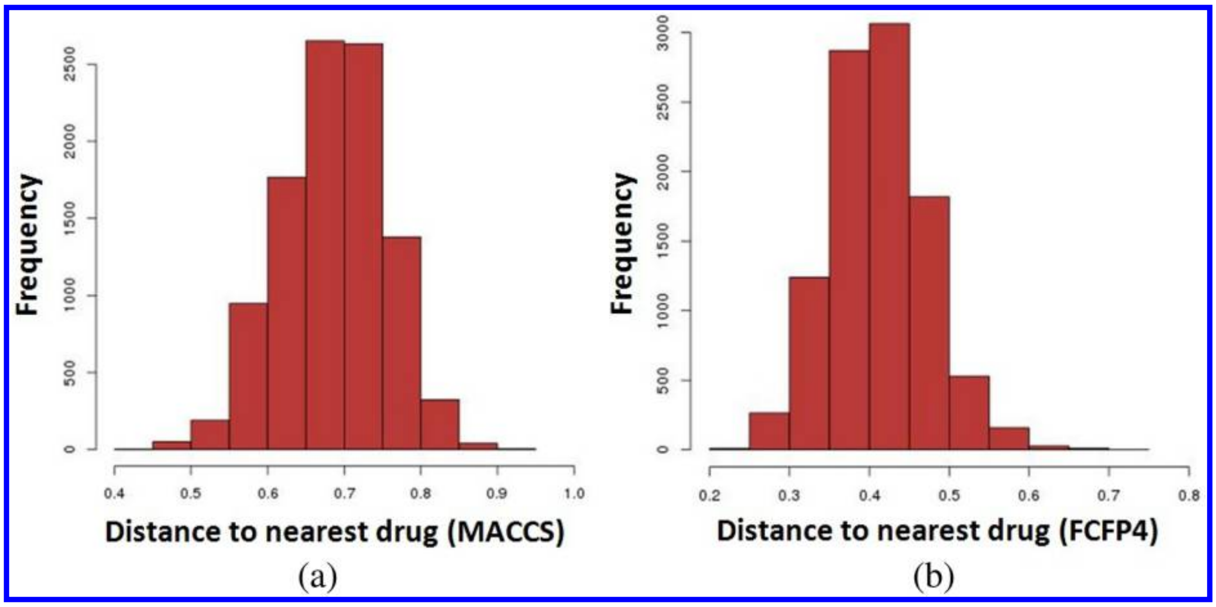


Figure 7. (a) Tanimoto score distributions of the 9994 products in the S sample compared with the 1540 compounds of DrugBank using (a) MACCS keys and (b) FCFP4 fingerprints.

within the entire population and within the S sample were of similar distribution, we plotted the reaction frequencies as a histogram (Figure 4). It is intriguing that almost 75% of the generated products of SCUBIDOO are based on four chemical reactions: Schotten–Baumann amide (33.3%), Buchwald–Hartwig (17.5%), reductive amination (13.6%), and Negishi (10.5%). This partitioning is of course related to the distribution of reagent classes observed in the previous section. It is also similar to the findings of Hartenfeller et al.[12] and lines up with the study of Roughley and Jordan,[18] where these four reactions are among the top six reactions used in the pharmaceutical field. In contrast, the popular Suzuki coupling is underrepresented here. This is due to the fact that only a few boronic acids (97) were present in the initial building block library.

Chemical Properties. In the second stage of stratified balanced sampling, the representative products of each stratum were selected using auxiliary variables (i.e., molecular descriptors). For the purpose of comparing SCUBIDOO to the S sample, the distributions of these auxiliary variables for each data set were plotted as histograms (Figure 5). The distributions of the two data sets are similar, indicating that the representative sample respects the heterogeneity of the initial population.

Diversity and Novelty. Analysis of the space spanned by the aforementioned physicochemical properties (Figure 6) using the two main principal components depicts the S sample as overlapping with the property regions of known drugs and lead-like compounds. This suggests that many of the generated products are in principle drug-like. However, despite this overlap, there are also many molecules from the S sample that are located within regions where known drugs and lead-like compounds are absent, indicating the existence of potentially chemically novel compounds. The composition of features of each plane is provided in Table S5 in the Supporting Information.

We also compared the S sample against DrugBank in more detail. For each product from the S sample, the distance to the nearest drug was calculated using MACCS keys[41] and FCFP4 fingerprints employing the Tanimoto coefficient. We used two different fingerprints for this comparison because we wanted to obtain two different opinions on similarity. The MACCS fingerprint contains 166 bits and is used for substructure searching. Each bit position specifically encodes a common functional group. In contrast, FCFP4 fingerprints are topological circular fingerprints, which are not predefined and can represent a large number of different molecular features. Therefore, these fingerprints highlight how particular features of known drugs are retrieved within SCUBIDOO products. Tanimoto score frequencies are plotted in Figure 7. Interestingly, the distribution for MACCS keys similarity is centered around a Tanimoto score of 0.7, suggesting that the S sample contains both similar and dissimilar products in comparison with known drugs. We treat 0.7 as a cutoff between similar and dissimilar compounds for MACCS keys, as it was shown to have discriminative power in an earlier study.[42] In the case of FCFP4 fingerprints, the distribution is centered around 0.4, which is also known as a good discriminative cutoff.[43] This maximum at a lower value is expected, as FCFP4 fingerprints are stricter in terms of Tanimoto score when dealing with bigger molecules. The fractions of dissimilar products obtained using two different fingerprints can be interpreted as hints that SCUBIDOO contains novel chemical entities in comparison with known drugs.

Synthetic Accessibility. To obtain a second computational assessment of the ease of synthesis of the products within SCUBIDOO, the SA score was computed for each product. The distribution of SA scores, plotted in Figure 8, is centered around of value of 3 with the vast majority (96%) lying below an SA score of 4, indicating easy-to-make products rather than overcomplex molecules.

Application: Retrospective Studies. In order to exemplify how SCUBIDOO might be useful in a ligand discovery context and to demonstrate several usage scenarios, retrospective similarity screening campaigns are presented here. The main goal is to show how SCUBIDOO can be used to retrieve known drugs or highly similar analogues. Table 1 lists all of the examples that were analyzed. The first example is meant to be comprehensive in order to demonstrate how to use SCUBIDOO step-by-step. Only examples 1, 2, and 3 are described here. The remaining examples are provided in the Supporting Information.

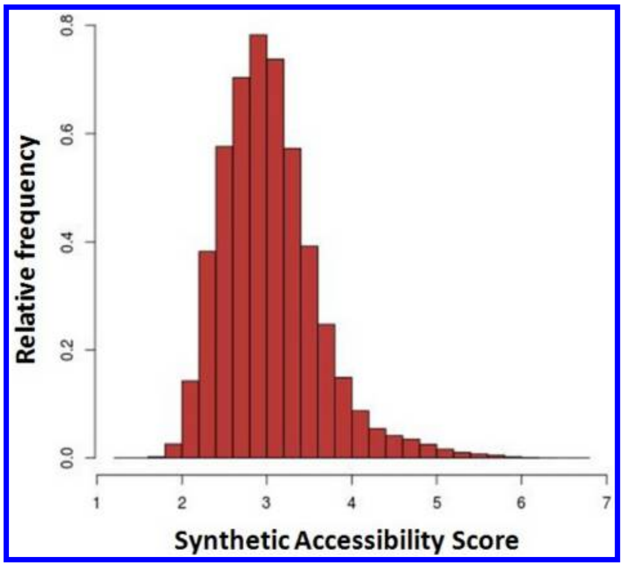


Figure 8. Distribution of SA scores for all of the products contained in SCUBIDOO.

Example 1: DB08235. Ligand-Based Strategy. In the similarity screen between DrugBank and the S sample, a close match was identified between the experimental drug DB08235 and SCUBIDOO product S10143065 (Figure 9), with an FCFP4 Tanimoto score of 0.69. Both molecules have an amide group attached to an indole moiety. DB08235 is an experimental drug that was identified as an inhibitor of the Arp2/3 complex[44] and may be utilized as potential anticancer agent. A *one-click* search in SCUBIDOO's Web interface retrieved information for S10143065. It is predicted to be synthesizable using a Schotten−Baumann amide reaction between the two building blocks 4029192 and 4089476 (Scheme 1). Building block 4089476 is particularly interesting here, as it contains the indole moiety. We then looked for every product in SCUBIDOO made from building block 4089476 using a Schotten−Baumann amide reaction, since this reaction introduces an amide bond, and 4460 derivative products were retrieved and compared to DB08235 using FCFP4 fingerprints. This search identified five products with Tanimoto scores above 0.84 (Figure 10). Among those, two very close analogues of DB08235 were present. Indeed, products S00003866 and S00021706 contain an isoxazole moiety and a thiazole moiety, respectively, which are very close to the thiophene moiety of DB08235. This application shows that after a molecule in the S sample that is similar to a given drug is identified, we can use the synthetic information on the product of interest to quickly screen the entire SCUBIDOO data set using a building block identifier and a reaction. A full search of the entire library would have taken several hours, as opposed to the few minutes for screening of the S sample and the 4460 derivatives of the original "hit". Therefore, we were able to efficiently analyze the analogues based on building block 4089476 and retrieve five products that are more similar to the drug than the initial hit found in the S sample. This example also illustrates how SCUBIDOO can be used for structure−activity relationship (SAR) studies or to generate suggestions for fragment-growing strategies.

Structure-Based Assessment of DB08235 Analogues. The five analogues of DB08235 were then docked into the Arp2/3

Table 1. Summary of the Hits Found within SCUBIDOO after Similarity Screening against DrugBank and the PDB Based on the FCFP4 Fingerprints

ex[a]	ref ID[b]	hit ID[c]	sim (FCFP4)[d]	reaction[e]	ref set[f]	sample[g]
1	DB08235	S00003866	0.96	amide	DrugBank	S
2	DB01097	S00131967	0.68	amide	DrugBank	M
3	H50	S02142952	0.93	Suzuki	PDB	M
3	H50	S02148982	0.94	Suzuki	PDB	M
4	4K6	S07366028	1	amide	PDB	S
5	F8E	S13393814	1	Buchwald−Hartwig	PDB	L
6	RM8	S01918821	1	amination	PDB	L
7	1DZ	S16929461	1	N-arylation	PDB	L

[a] Example. [b] Reference ID (drug or PDB ligand). [c] SCUBIDOO ID. [d] Tanimoto similarity score using the FCFP4 fingerprints. [e] Predicted reaction. [f] Reference data set. [g] SCUBIDOO sample.

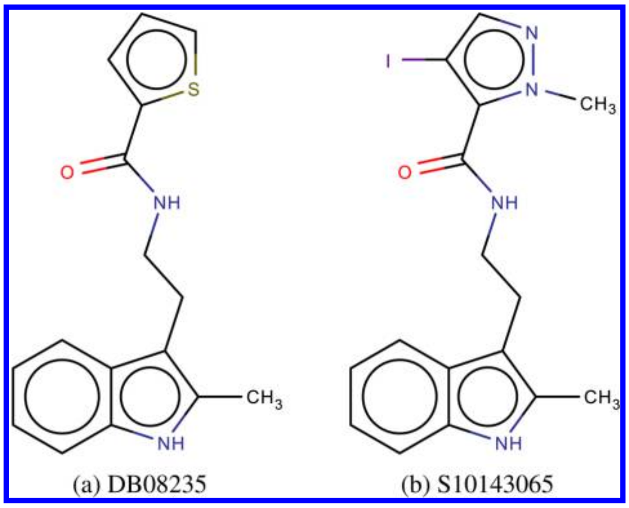


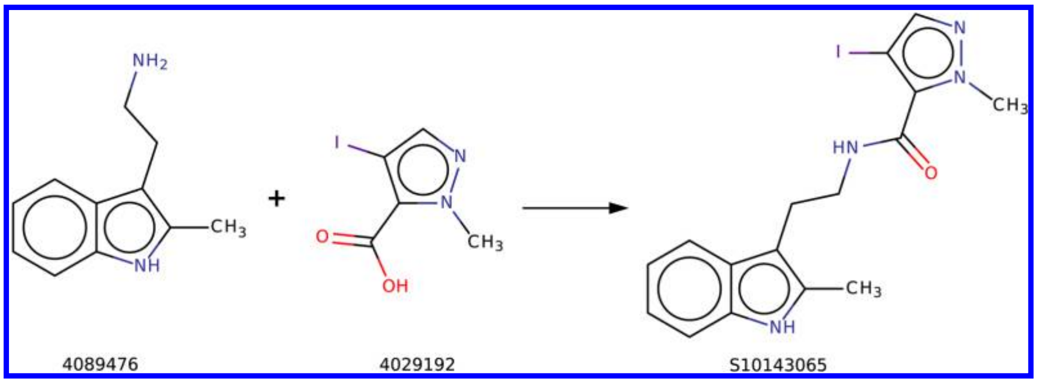
Figure 9. Molecules mentioned in example 1: (a) DrugBank compound DB08235 and (b) its closest similar product in the S sample, S10143065.

complex using FRED.[45] The crystal structure is available in the PDB (ID 3DXK). The predicted binding mode for product S00021706, illustrated in Figure 11, overlaps perfectly with the cocrystallized ligand DB08235. The same binding mode was also retrieved for the remaining four analogue products. This can be taken as a hint that the two close analogues might exhibit similar biological activities, as they scored favorably in this orthogonal screening method. Another way to look at this result is that if S00021706 had been suggested as a potential ligand in an unbiased docking screen, one would have been able to readily retrieve the other analogues quickly, leading to potentially biologically active compounds.

Example 2: DB01097. After screening of the DrugBank against the M sample, a close match was identified between leflunomide (DB01097) and product S00134656 (Figure 12) with an FCFP4 Tanimoto score of 0.67. They both contain building block 3001678. Focusing on this building block, similar to the procedure of example 1, led to the identification of the closest product to leflunomide in the entire SCUBIDOO database, S00131967, with an FCFP4 Tanimoto score of 0.68. The only difference between S00131967 and leflunomide is that the benzene ring is replaced by a pyridine ring. However, the isoxazole ring, which is the active part and is opened upon administration,[46] is identical. S00131967 is predicted to be synthesizable using a Schotten−Baumann amide reaction. This reaction was also applied to synthesize leflunomide.[47]

Example 3: H50. In the comparison of the ligands from the PDB against the M sample, a similarity appeared between the fibrillogenesis inhibitor H50[48] (Figure 13a) and product S03544112 with an FCFP4 Tanimoto score of 0.84 (Figure 13b). Product S03544112 is predicted to be synthesizable using a Suzuki coupling between the building blocks 4003301 and 6644827. Suzuki coupling was also applied to synthesize H50.[49] In this case, the two molecules do not share a common building block. Exploring the analogues of building block 4003301 led to the boronic-acid-containing building block 3200974, which is more similar to the initial H50, as only a chlorine is replaced by a fluorine. The derivatives of building block 3200974 obtainable by Suzuki coupling were then compared to H50, and product S02142952 (Figure 13c) was identified as a closer analogue, with an FCFP4 Tanimoto similarity score of 0.93. Furthermore,

Scheme 1. Route for Obtaining S10143065 (right): The Schotten−Baumann Amide Reaction between the Two Building Blocks 4089476 (left) and 4029192 (middle)



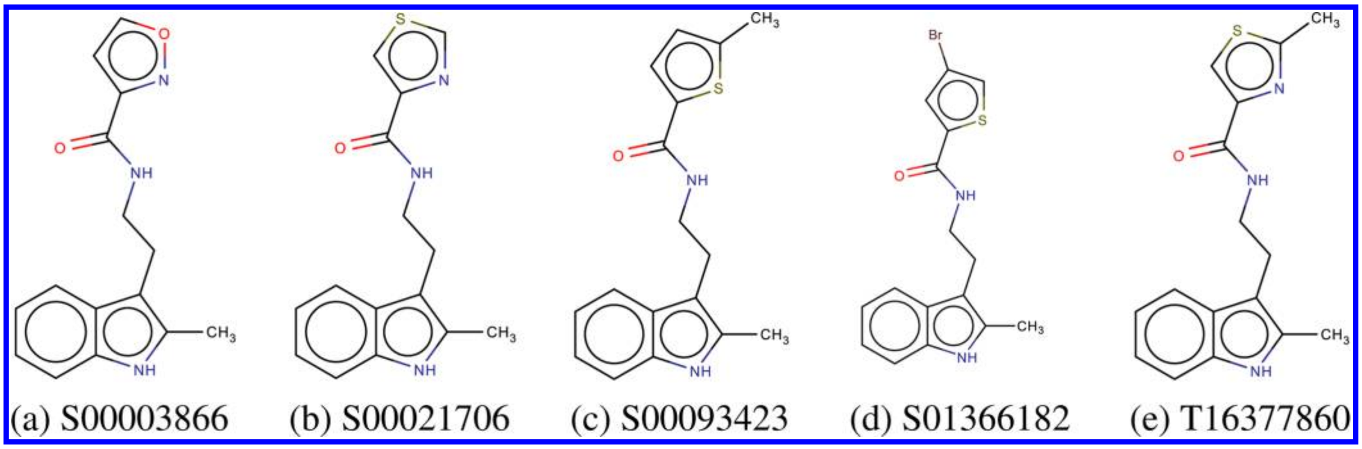


Figure 10. (a−e) Molecular analogues of DrugBank compound DB08235 found in SCUBIDOO.

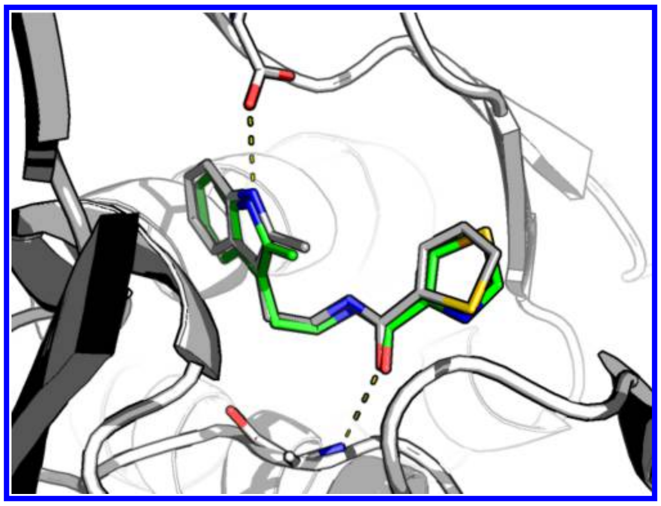


Figure 11. Docking: predicted binding mode of product S00021706 (green carbons) compared to the crystallized ligand DB08235 (gray carbons). The protein is shown in white cartoon representation, and H-bonds are indicated by yellow dashed lines.

the derivatives of building block 4003301 were compared to H50, leading to yet another close analogue, product S02148982 (Figure 13d), with a similarity score of 0.94.

■ DISCUSSION AND CONCLUSIONS

We have presented a freely accessible database concept currently holding 21 million screenable chemical products, each coming with synthesis information allowing an estimation of how readily it might be obtained. All of the reactants used for the creation of this database are publicly available, and the reactions employed are among the most popular ones in the pharmaceutical field. The products are accessible using an intuitive Web interface, complete with synthesis instructions. SCUBIDOO is unique because it not only provides reaction suggestions but also clearly specifies potential side or alternative reactions and even reactive groups that could interfere during synthesis. Such warnings can help chemists as decision tools during synthesis planning and protecting group design. Moreover, since SCUBIDOO is a Web-based application, gathering of feedback from the scientific community is possible and will be applied to refine the products and reactions in future versions. This will benefit the community by making the set of reactions and the alerts in SCUBIDOO even more robust.[17]

It is clear that the diversity of the initial building block library has a large impact on the generated products. Therefore, future versions or derivative libraries will originate from new building blocks and increase the diversity. Conversely, in order to broaden the library more, it could be interesting to ensure that the building block library provides some heterogeneity within the reagent classes. This will enhance the diversity of the generated products and allow the use of the set of 58 reactions at its full potential. As a possible future development of their work,[23] Goldberg et al. suggested an open-innovation approach where ideas for novel structures to synthesize could be accessed from external sources. SCUBIDOO fits right into this context.

SCUBIDOO is not so much a database as it is a concept of how to enter chemical space. Along those lines, the PCA analysis suggested a share of novelties within SCUBIDOO that can be assumed to be “low-hanging fruits” as described by Hartenfeller et al.[12] More precisely, such chemical scaffolds undescribed for a particular target can be used as starting points

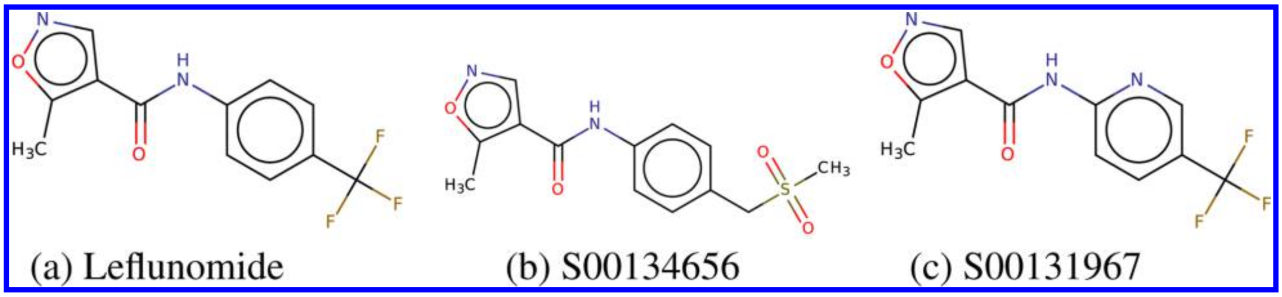


Figure 12. (a) DrugBank compound DB01097 (leflunomide). (b) Its closest similar product in the S sample, S00134656 (Tanimoto score = 0.67). (c) Its closest similar product in all of SCUBIDOO, S00131967 (Tanimoto score = 0.68).

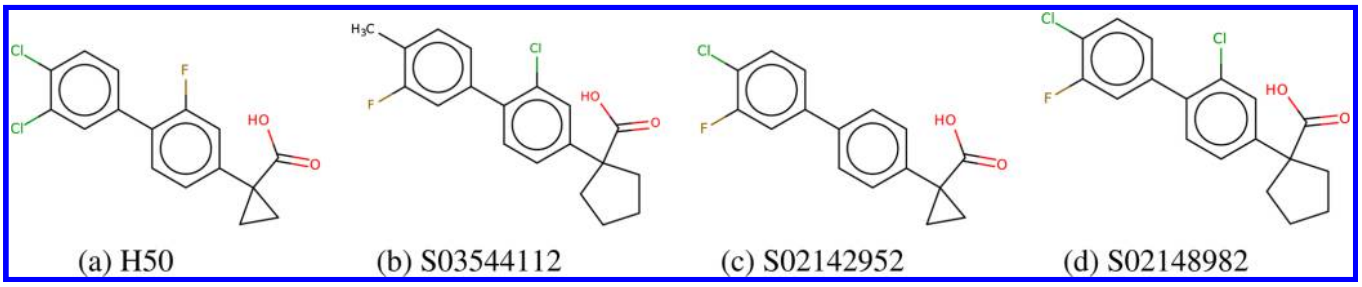


Figure 13. (a) Ligand H50. (b) Its closest similar product in the M sample, S03544112 (Tanimoto score = 0.84). (c) A similar product in SCUBIDOO, S02142952 (Tanimoto score = 0.93). (d) Its closest similar product in SCUBIDOO, S02148982 (Tanimoto score = 0.94).

for ligand discovery projects. Once such a "fruit" is identified, SCUBIDOO allows users to rapidly explore the tree around this point in order to harvest close analogues. This analogy is illustrated in the retrospective studies, where we identified existing active molecules within SCUBIDOO starting from the samples. Whenever possible, we retrieved information about the original synthesis, which matches the category of SCUBIDOO well. In a prospective screening setting, where such suggestions might come from docking, SCUBIDOO can be used in a straightforward manner to assemble a tailored library. Since this method also proceeds via a close look at the fragment composition, it fits right in with the ALTA approach, which we have described earlier.[50] Additionally, in cases where only analogues of active molecules were retrieved within the samples, a second, refined, search at the building block level allowed us to retrieve the initial active molecule or a very close analogue in all cases discussed here. Retrospective binding mode analysis showed that the small chemical differences should not affect the ligand−protein interactions.

While the size of this database is relatively small compared with those of other virtual chemical space libraries,[15] navigating through 21 million entities already presents a challenge. In order to provide an entry point, SCUBIDOO was reduced to three different representative samples denoted as S, M, and L, containing 9994, 99 977, and 999 794 compounds, respectively. The representative samples were extracted using a stratified balanced sampling algorithm, which is a well-known algorithm for population surveys. Its application to a large molecule set allowed us to obtain samples respecting the heterogeneity of the initial set. This is essential in order to use the database in an efficient fashion in future screens. We are aware that 21 million products is far from covering chemical space. Nevertheless, we think that concomitant with future expansion of our database, such reduction steps need to be applied in order to keep the data computationally tractable. To the best of our knowledge, this is the first application of balanced sampling for this purpose.

A typical application protocol might start with screening of the SCUBIDOO sample of a user's choice using a ligand-based strategy, a structure-based strategy, or both. The second step would then involve a focused search around the candidate hits found during the first screening, this time extracting molecules from the entire database.

Furthermore, SCUBIDOO can also be employed as a growing strategy within a fragment-based project or as an SAR tool. Indeed, any SCUBIDOO product is the assembly of two building blocks (or fragments). If one of those building blocks shows promising interactions with a given target, SCUBIDOO lets users quickly retrieve all of the derivatives of this building block. All of these products can then be downloaded in SMILES format for further investigation.

While the retrospective assessment shows the existence of known active molecules within SCUBIDOO and the SA scores suggest that most of the products fall on the side of relatively facile synthetic realization, the next step will be to validate products experimentally. We think that this database has the potential to go in the direction of one of the expected breakthroughs in future de novo design, as stated by Schneider:[51] "*reliable prediction of the synthesizability of new chemical entities and suggestion of short synthesis routes and reactions, directly coupled to integrated synthesis-and-test platforms*". After the recent publication of the "synthesis machine",[5] it is not ludicrous to think that the first brick of such a workflow might be a database based on the SCUBIDOO concept.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00203.

Spreadsheet containing the library of 58 reactions (encoded in SMARTS notation), the reagent classes, the nucleophile and electrophile groups (encoded as SMARTS), the 13 reactions that do not appear in SCUBIDOO, and PCA results on diversity and novelty (XLSX)

Discussion of retrospective study examples 4−7 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: peter.kolb@uni-marburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

P.K. thanks the German Research Foundation (DFG) for Emmy Noether Fellowship KO4095/1-1 and the LOEWE Program for participation in the research cluster SynChemBio. The authors acknowledge J. Cramer, Dr. J. Schiebel, Dr. A. Metz, and C. Taylor for their valuable input through discussions and reading of the manuscript. The authors also thank OpenEye Scientific Software for providing the software referenced in this study and permitting the generated products to be made publicly available.

■ REFERENCES

(1) Lederberg, J. Topological mapping of organic molecules. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *53*, 134−139.

(2) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discove. *J. Chem. Inf. Model.* **2007**, *47*, 342−353.

(3) Reutlinger, M.; Rodrigues, T.; Schneider, P.; Schneider, G. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angew. Chem., Int. Ed.* **2014**, *53*, 4244−4248.

(4) Service, R. F. The Synthesis Machine. *Science* **2015**, *347*, 1190−1193.

(5) Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Synthesis of many different types of organic small molecules using one automated process. *Science* **2015**, *347*, 1221−1226.

(6) Alvim-Gaston, M.; Grese, T.; Mahoui, A.; Palkowitz, A. D.; Pineiro-Nunez, M.; Watson, I. Open Innovation Drug Discovery (OIDD): a potential path to novel therapeutic chemical space. *Curr. Top. Med. Chem.* **2014**, *14*, 294−303.

(7) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(8) Kolb, P.; Irwin, J. J. Docking screens: right for the right reasons? *Curr. Top. Med. Chem.* **2009**, *9*, 755−770.

(9) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D.; Khachko, D.; Fokin, V.; Queen, C.; Zosimov, V. A very large diversity space of synthetically accessible compounds for use with drug design programs. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 47−63.

(10) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and searching 1020 synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 341−350.

(11) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497−520.

(12) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S. Probing the bioactivity-relevant chemical space of robust reactions and common molecular building blocks. *J. Chem. Inf. Model.* **2012**, *52*, 1167−1178.

(13) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching fragment spaces with feature trees. *J. Chem. Inf. Model.* **2009**, *49*, 270−279.

(14) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. Methods in Molecular Biology. *Chemical Library Design* **2011**, *685*, 253−276.

(15) Peng, Z. Very large virtual compound spaces: Construction, storage and utility in drug discovery. *Drug Discovery Today: Technol.* **2013**, *10*, e387−e394.

(16) Gasteiger, J. Cheminformatics: Computing Target Complexity. *Nat. Chem.* **2015**, *7*, 619−620.

(17) 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*, 3093−3098.

(18) 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.

(19) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011**, *2*, 349−355.

(20) ChemBridge. Building Blocks. http://www.chembridge.com/building_blocks/.

(21) Landrum, G. RDKit: Open-source chemoinformatics. http://www.rdkit.org.

(22) Cecchini, M.; Kolb, P.; Majeux, N.; Caflisch, A. Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J. Comput. Chem.* **2004**, *25*, 412−422.

(23) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discovery Today* **2015**, *20*, 11−17.

(24) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(25) *OMEGA*, version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM; http://www.eyesopen.com.

(26) Chauvet, G. Stratified balanced sampling. *Surv. Methodol.* **2009**, *35*, 115−119.

(27) Grafstrom, A. BalancedSampling: Balanced and spatially balanced sampling. In R package version 1.4, 2014.

(28) *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(29) Deville, J.-c.; Tillé, Y. Efficient balanced sampling: The cube method. *Biometrika* **2004**, *91*, 893−912.

(30) Hasler, C.; Tillé, Y. Fast balanced sampling for highly stratified population. *Comput. Stat. Data Anal.* **2014**, *74*, 81−94.

(31) DrugBank: Open Data Drug & Drug Target Database, version 4.1; http://www.drugbank.ca/.

(32) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(34) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(35) *QUACPAC*, version 1.6.3.1; OpenEye Scientific Software: Santa Fe, NM; http://www.eyesopen.com.

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

(37) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(38) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(39) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Model.* **1999**, *39*, 868−873.

(40) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(41) McGregor, M. J.; Pallai, P. V. Clustering of Large Database of Compounds: Using MDL keys As Structural Descriptors. *J. Chem. Inf. Model.* **1997**, *37*, 443−448.

(42) Chevillard, F.; Lagorce, D.; Reynès, C.; Villoutreix, B. O.; Vayer, P.; Miteva, M. Multimodel Protocol Based on Chemical Similarity In silico Prediction of Aqueous Solubility: A Multimodel Protocol Based on Chemical Similarity. *Mol. Pharmaceutics* **2012**, *9*, 3127−3135.

(43) Wawer, M.; Bajorath, J. Similarity-potency trees: A method to search for SAR information in compound data sets and derive SAR rules. *J. Chem. Inf. Model.* **2010**, *50*, 1395−1409.

(44) Nolen, B. J.; Tomasevic, N.; Russell, A.; Pierce, D. W.; Jia, Z.; McCormick, C. D.; Hartman, J.; Sakowicz, R.; Pollard, T. D. Characterization of two classes of small molecule inhibitors of Arp2/3 complex. *Nature* **2009**, *460*, 1031−1034.

(45) Mcgann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578−596.

(46) Liu, S.; Neidhardt, E.; Grossman, T.; Ocain, T.; Clardy, J. Structures of human dihydroorotate dehydrogenase in complex with antiproliferative agents. *Structure* **2000**, *8*, 25−33.

(47) Ivashkin, P.; Lemonnier, G.; Cousin, J.; Grégoire, V.; Labar, D.; Jubault, P.; Pannecoucke, X. $[^{18}\mathrm{F}]\mathrm{CuCF}_3$: A $[^{18}\mathrm{F}]$Trifluoromethylating Agent for Arylboronic Acids and Aryl Iodides. *Chem. - Eur. J.* **2014**, *20*, 9514−9518.

(48) Zanotti, G.; Cendron, L.; Folli, C.; Florio, P.; Imbimbo, B. P.; Berni, R. Structural evidence for native state stabilization of a conformationally labile amyloidogenic transthyretin variant by fibrillogenesis inhibitors. *FEBS Lett.* **2013**, *587*, 2325−2331.

(49) Peretto, I.; et al. Synthesis and biological activity of flurbiprofen analogues as selective inhibitors of beta-amyloid42 secretion. *J. Med. Chem.* **2005**, *48*, 5705−20.

(50) Kolb, P.; Berset Kipouros, C.; Huang, D.; Caflisch, A. Structure-based tailoring of compound libraries for high-throughput screening: Discovery of novel EphB4 kinase inhibitors. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 11−18.

(51) Schneider, G. Future de novo drug design. *Mol. Inf.* **2014**, *33*, 397−402.

DOI: 10.1021/acs.jcim.5b00203
J. Chem. Inf. Model. 2015, 55, 1824−1835