JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

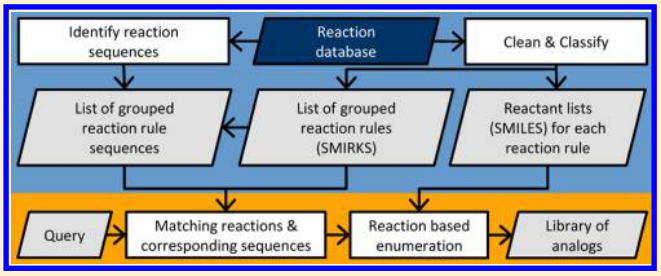
pubs.acs.org/jcim

Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration

Clara D. Christ,*[,‡] Matthias Zentgraf,[‡] and Jan M. Kriegl[‡]

[‡]Department of Lead Identification and Optimization Support, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorferstrasse 65, 88397 Biberach an der Riss, Germany

ABSTRACT: An approach to automatically analyze and use the knowledge contained in electronic laboratory notebooks (ELNs) has been developed. Reactions were reduced to their reactive center and converted to a string representation (SMIRKS) which formed the basis for reaction classification and *in silico* (retro-)synthesis. Of the SMIRKS that occurred at least five times, 98% successfully regenerated the original product. The extracted reaction rules (SMIRKS) and corresponding reactants span a virtual chemical space which showed a strong dependence on the size of the reactive center. Whereas relatively few robust reaction types were sufficient to describe a large part of all reactions, considerably more reaction rules were necessary to cover all reactions in the ELN. Furthermore, reaction sequences were extracted to identify frequent combinations and diversifying reaction steps. Based on the extracted knowledge a (retro-)synthesis tool was built allowing for *de novo* design of compounds which have a high chance of being synthetically accessible. In an example application of the *de novo* design tool, various feasible retrosynthetic routes to the query molecule were obtained. Reaction based enumeration along the top ranked route yielded a library of 29 920 compounds with diverse properties, 99.9% of which are novel in the sense that they are unknown to the public domain.



1. INTRODUCTION

Electronic laboratory notebooks (ELNs) nowadays used by medicinal chemists in the pharmaceutical industry are not only of documentary value but constitute a goldmine waiting to be exploited. They contain the answer to questions such as "How often do we perform which types of reactions and what is the average yield?" or, "Where in a multi-step synthesis is diversity introduced?". Furthermore, an ELN opens up a large chemical space of virtual molecules accessible by established chemistry, namely the chemistry contained in it. That is, the ELN contains all the information needed for *de novo* design[1−3] of virtual compounds which are by construction synthetically accessible.

Limited synthetic accessibility of *de novo* designed compounds has hampered the routine use of this technique in the pharmaceutical industry. Molecules may be scored for synthetic complexity after virtual generation,[4−7] or the building process may be adapted to bias the search toward synthetically accessible compounds. In the FOG method[8] fragments are connected based on fragment-fragment connection frequencies obtained from a training database. Infrequent connections are, therefore, avoided, and synthetic accessibility of the generated molecules is increased. In this way synthetic accessibility is incorporated into the building process without teaching the computer explicitly about chemical reactions. *In silico* generation of molecules can, however, also directly follow synthetic routes. The underlying reaction database can either be manually assembled or automatically derived from existing reaction databases. The SYNOPSIS[9] program uses a set of 70 hand-coded reaction types which are applied to starting molecules to generate new molecules. Estimates of reactivity are implemented to judge whether a reaction can be applied to a certain molecule or not. Hartenfeller et al.[10] have recently assembled a set of reactions useful for *de novo* design encoded as Reaction SMARTS[11] which focuses on cyclization reactions. A *de novo* design approach which uses reaction rules automatically derived from reaction databases has been presented by Patel et al.[12,13] Reaction rules are encoded as reaction vectors[14] which represent "the structural changes that take place at the reaction center along with the environment in which the reaction occurs".[12]

Searching in virtual fragment spaces is another way of obtaining synthetically accessible new chemical entities.[15−18] One approach is to encode established combinatorial chemistry protocols in machine readable form. Rather than enumerating the prohibitively large virtual combinatorial space, it is stored as a set of core molecules with corresponding reagent lists containing suitable R-group molecules as well as rules how to connect cores and R-groups.[15,17,18] At Boehringer Ingelheim the whole in-house knowledge on combinatorial chemistry has been made searchable in form of a large virtual library called BICLAIM (Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules).[18] Special search techniques which avoid explicit enumeration have been developed to efficiently search these virtual combinatorial spaces.[17,18]

Received: March 2, 2012

dx.doi.org/10.1021/ci300116p | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

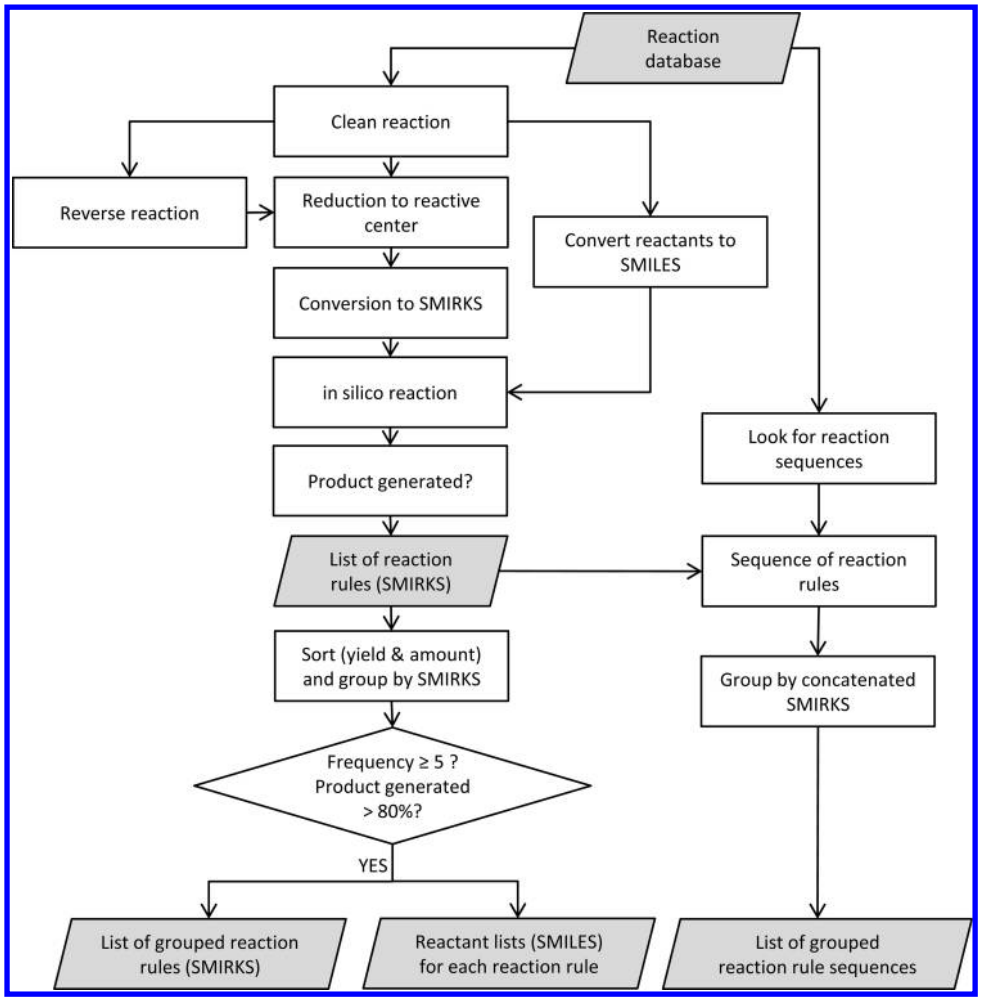


Figure 1. Simplified depiction of the various steps involved in reaction and reaction sequence processing (see section 2 for details).

Rather than focusing on molecule generation, computer-aided synthetic design (CASD) focuses on generating synthetic routes. [For an overview of the different approaches see ref 19 and Figure 1 of ref 20.] As in *de novo* design methods, the knowledge base of reaction rules employed may be manually assembled or stem from reaction databases. Automated extraction and machine learning approaches have been used in CASD for about 20 years.[20] A recent method that assembles its knowledge base in an automated fashion is "Route Designer".[20] Reaction cores are automatically extracted and then extended in order to capture the relevant neighboring functional groups that have an impact on reactivity. Although the core extraction is automatic, many postprocessing steps such as the core extension use sophisticated hand-coded rules.

Here, we present an approach that allows one to automatically analyze and use the knowledge contained in ELNs. Reactions are reduced to their reactive center enabling classification and subsequent reaction type based analyses. These analyses include calculating mean properties such as average yield or judging the potential for diversification of a certain reaction type. Subsequently, reaction sequences are analyzed to identify frequent reaction combinations and to determine the interesting, i.e., diversifying steps of a reaction sequence. Finally, the extracted knowledge of reactions and reaction sequences is used to build a retrosynthesis and reaction based enumeration program allowing for *de novo* design of synthetically accessible compounds.

2. MATERIALS AND METHODS

Processing of reactions consisted of several steps including reaction cleaning, reduction to the reactive center, conversion to SMIRKS, and classification. These steps are described in sections 2.1−2.5, whereas section 2.6 discusses reaction sequence processing. The complete knowledge extraction process is summarized in Figure 1. How this knowledge is then used to build an *in silico* (retro-)synthesis tool is described in section 2.7 and depicted in Figure 3.

2.1. Data Source and Software Tools. The basis for this study consisted of a set of 402 546 medicinal chemistry reactions taken from our in-house electronic laboratory notebook (ELN). About 4% of the reactions belong to a combinatorial chemistry library. All programs for reaction processing and virtual (retro-)synthesis were written in C++ using the OEChem toolkit[21] and OCCI.[22] Postprocessing was performed using KNIME.[23] Pipeline pilot[24] reporting facilities were used to build a user-friendly front-end for the retrosynthesis and enumeration tool described below.

2.2. Reaction Cleaning. For each reaction the reaction description in Rxnfile-format[25] including atom map indices as well as associated information was retrieved from the database. Reactions without information about product yield and amount were discarded. We chose to keep only the maximum yield product and deleted any further products from the rxn-entry. The atom map indices, i.e., indices mapping a reactant atom to

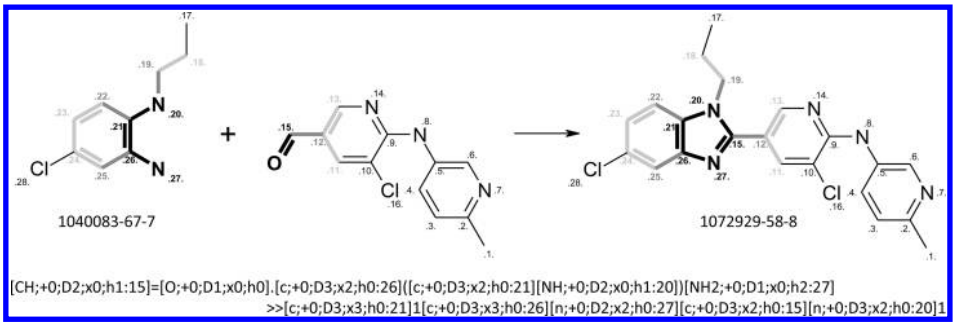


Figure 2. Cropping the reaction to its reactive center. Black bold lines and atom names indicate the smallest reactive center (*small*) which consists solely of atoms for which properties differ on reactant and product side. Including first neighbors, the medium sized reactive center is obtained (*medium*, bold dark gray and black). The large reactive center is obtained by including also second neighbors (*large*, bold light gray, dark gray, and black). The SMIRKS string derived from the *small* reactive center (bold black) is also shown. (Note that this example reaction is not part of the investigated ELN; product taken from ref 29. CAS Registry numbers are given.)

a product atom, were used for further reaction cleaning. As atom map indices are assigned automatically by the ELN and are not manually curated, erroneous mappings may occur. Complete molecules without atom map indices were deleted from the reaction. Subsequently atom map indices were checked for correctness. In some cases a reactant atom map index occurred multiple times on the product side. This may happen if there are multiple products or if the same functional group is introduced several times into the product. In the former case the reaction would be split into separate reactions to resolve the multiply occurring atom map indices. As we decided to keep only the maximum yield product we did not encounter this; however, the algorithm would be able to handle these cases. The latter case is resolved by multiplying the corresponding reactant and adjusting atom map indices on reactant and product side such that each atom map index occurs only once on reactant and product side. No further molecule standardization was performed as reactant and product molecules are standardized upon registration into the database.

2.3. Reduction to the Reactive Center. For reaction classification and later described applications such as reaction based enumeration, the essential reaction information needed to be extracted. To this end reactions were reduced to their reactive center as depicted in Figure 2. For each reaction, reactive centers of three different sizes were derived. The smallest, least specific reactive center consisted only of pairs of atoms with differing properties (*small*). OEChem atom properties considered were as follows: the atom map indices of the neighboring atoms, the formal charge, the total hydrogen count, atomic hybridization, atom heavy degree, and the number of ring bonds. All atoms which only occurred on one of the sides of the reaction were also included. Including also first neighbors a medium sized reactive center was obtained (*medium*). Going up to second neighbors yielded the large reactive center (*large*). Technically, reduced reactions were obtained by deleting all atoms which were not part of the reactive center from the OEChem reaction object. The reaction center extension to one and two spheres of neighboring atoms is similar to other shell based approaches such as, e.g., the one used by InfoChem's reaction classification program ICClassify.[26] The larger the reactive center the better the description of the chemical environment. Note, however, that the influence of distant functional groups may be missed even with the *large* reactive center description. Similarly, neighboring atoms which are not chemically relevant may be included in the larger reactive centers. Rule based core extension algorithms such as the one implemented in "Route Designer"[20] allow extension to the relevant neighboring functionality only. However, as the rules encode specific structural features they are less generally applicable as the simple shell based approaches.

2.4. Conversion to SMIRKS. In order to use the reduced reaction description for applications such as reaction based library enumeration the OEChem reaction object needed to be exported in a reaction transform format. Common formats include MDL reaction query files (Rxnfiles)[25] and the Daylight reaction transform language SMIRKS.[11] As neither of these formats can be readily exported using OEChem we chose to convert the reaction object into reaction SMILES and modify this SMILES on a string basis to obtain a SMIRKS reaction description. The generated canonical SMILES contained atom map indices and the charge of the atom. We decided not to include atom and bond stereo information. For performance reasons we chose to handle hydrogens implicitly although this violates the strict Daylight SMIRKS definition. Encoded atom properties were the number of explicit connections (D < *n* >), the number of ring bonds (x < *n* >), and the implicit hydrogen count (h < *n* >). Furthermore, neutral atoms were explicitly marked as being neutral by adding a "+0" feature. Smarts features were added to all atoms in line with the OEChem SMIRKS implementation. An example *small* SMIRKS is shown in Figure 2.

The *large* SMIRKS descriptions, i.e. the SMIRKS corresponding to the large reactive center, were subjected to further tests in order to confirm their suitability for reaction based enumeration and *in silico* retrosynthesis. To check its integrity the SMIRKS was reparsed into an OEChem reaction object. The reactants and products were converted to SMILES and also reparsed. In a next step the reaction was applied to the reactants. Regeneration of the original product is a necessary test of the SMIRKS to be useful for reaction based enumeration. The suitability for *in silico* retrosynthesis was tested by repeating the procedure with a reversed reaction object. Here, regeneration of the original reactant was necessary for the SMIRKS to be suited.

2.5. Reaction Classification and Postprocessing. All reactions were sorted by yield category (70−100%, 30−70%, 0−30%, > 100%) and decreasing product amount. The following procedure was performed for each of the three reaction center sizes (*small*, *medium*, and *large*) separately. All reactions with the same SMIRKS string (neglecting atom map indices) were grouped together. The first reaction of each

dx.doi.org/10.1021/ci300116p | *J. Chem. Inf. Model.* XXXX, XXX, XXX−XXX

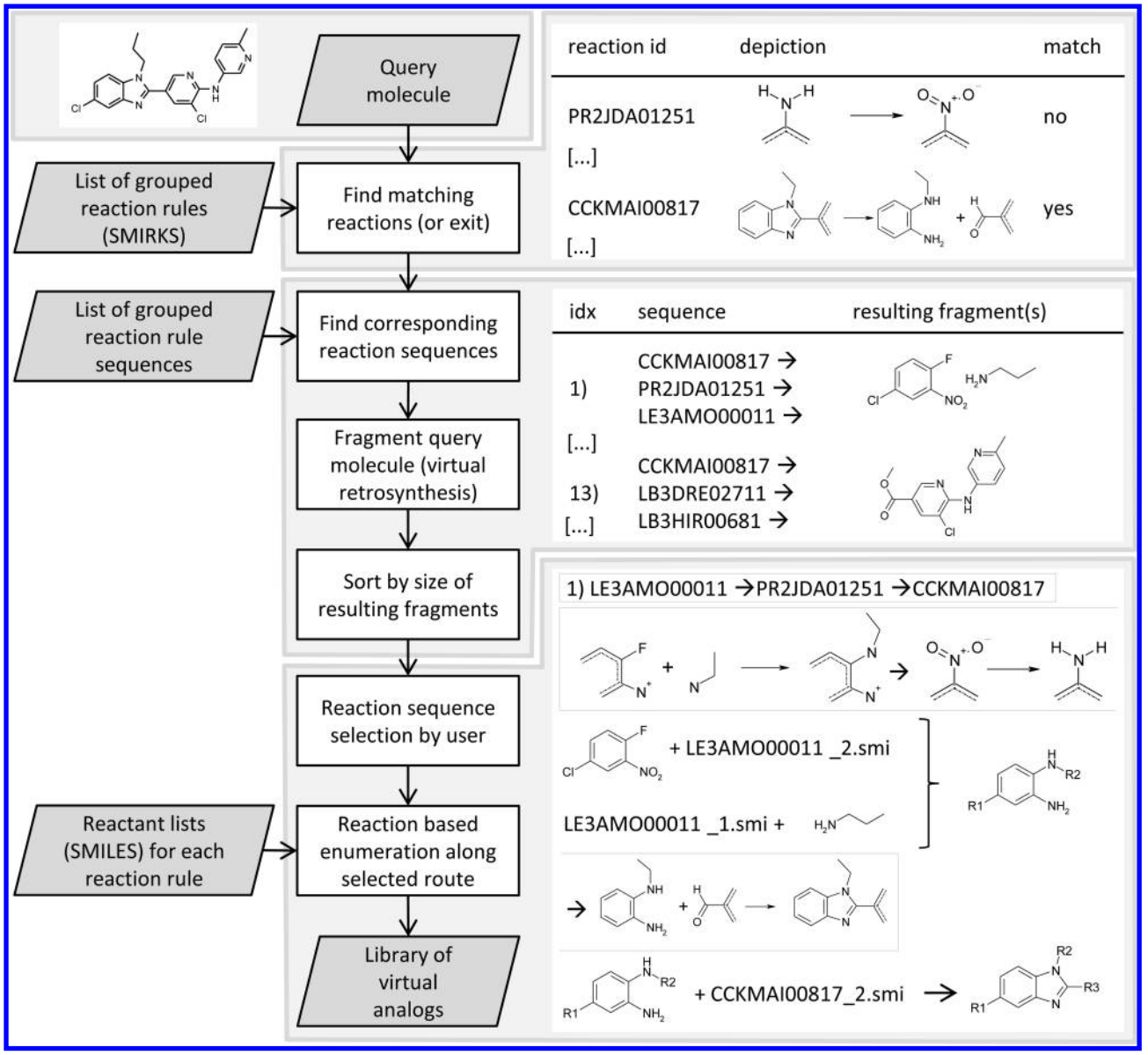


Figure 3. Simplified, schematic depiction of the various steps involved in the retrosynthesis and reaction based enumeration tool (see section 2.7 for details). Given a user defined query molecule a library of virtual analogs is obtained. The right-hand side exemplifies the different steps for the depicted query molecule. Note that the reaction schemes are simplified depictions which do not show all atom properties encoded in the SMIRKS. How the reaction rule and reactant files are obtained is shown in Figure 1 and detailed in sections 2.2−2.6.

group, i.e. according to the sorting a high yield and product amount example, was chosen as representative reaction experiment of the group. Each reaction type group was assigned the reaction type name most frequently used within the group (i.e., a consolidated reaction type name). The group size was determined and the average yield for this reaction type calculated. Reaction type groups were then sorted by group size, i.e. frequency of the reaction type. For each reaction type the number of distinct molecules used as reactant 1, reactant 2, and so on was determined to allow for a rough estimate of the number of possible virtual products. To judge the potential for diversification of a certain reaction type we calculated a variability score defined as

$$\text{variability score} = 1 - 1\Big/\frac{\text{Estimated number of virtual products}}{\text{Number of unique reactions of this type}} = 1 - 1\Big/\frac{\prod_{i=1}^{n}\text{Number of distinct molecules used as reactant } i}{\text{Number of unique reactions of this type}} \quad (1)$$

The score ranges from 0 to 1 with 0 indicating a low potential for diversification (e.g., a functional group transformation) and 1 indicating a high potential for diversification (e.g., an amide bond formation). For analysis purposes (see section 3.1 and

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

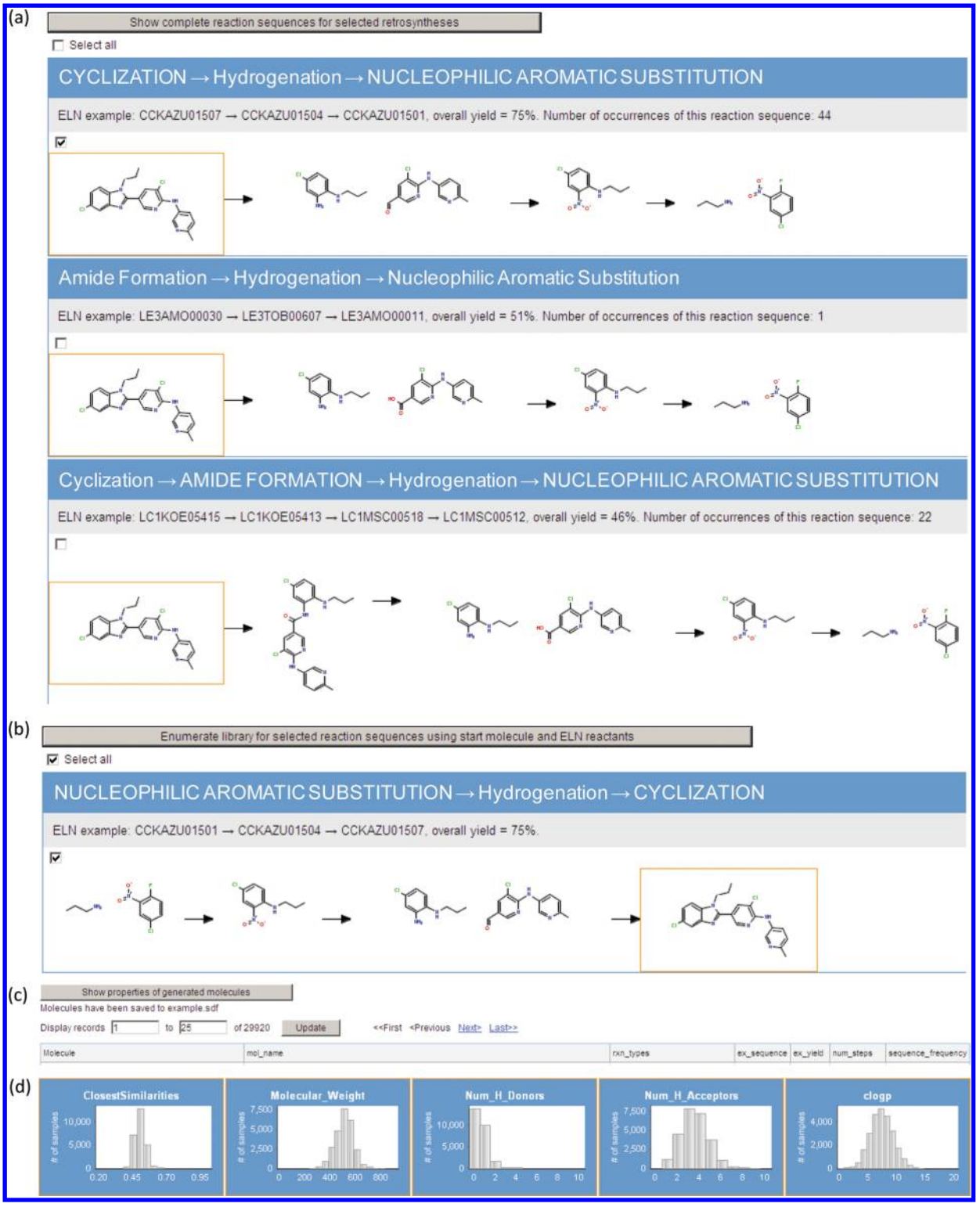


Figure 4. Screenshots showing the various steps of the retrosynthesis and reaction based enumeration tool from the user perspective. (a) Various virtual retrosynthetic routes for the query compound are suggested (note that only the first three are shown). Overall yield and frequency of related ELN syntheses which follow equivalent chemistry are given as an aid in the selection of the virtual retrosynthetic scheme to be applied. In this example, suggested sequences one and three should be more reliable than two as sequences of those types have been used more frequently. Capital font indicates a high diversification potential (see eq 1). (b) For the selected virtual retrosynthetic scheme(s) the complete reaction sequences are shown. In this example the sequence ends with the generation of the query molecule. In other examples the reaction sequence includes subsequent reactions going beyond query molecule formation. Clicking the submit button triggers enumeration of a library along the selected synthetic route. (c) A paged table shows the resulting molecules (not shown). Information includes a molecule name which allows the user to track the employed reagents. The user can then trigger the calculation of molecular properties of the enumerated library. (d) Calculated properties: Similarity to query molecule, molecular weight, number of H-bond donors and acceptors, and logP.

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Figure 7) a second grouping according to the consolidated reaction type name was performed.

Also in this processing step, the *large* SMIRKS patterns were subjected to further tests in order to be usable for reaction based library enumeration. As described above, we grouped by the SMIRKS string without atom maps. However, a SMIRKS including atom map indices extracted from the representative reaction experiment was chosen for later use. In addition to the procedure described above, a filter was applied passing only reaction types (SMIRKS) that occurred at least five times. Furthermore, we required that the original product was obtained in more than 80% of the cases when applying the SMIRKS to the original reactants. The same procedure was applied to the retrosynthetic *large* SMIRKS patterns. Only SMIRKS which passed the criteria in the synthetic and retrosynthetic direction were further pursued. Furthermore, for each synthetic reaction type (SMIRKS) all distinct reactants were sorted according to how often they were used and written to a file to make them accessible for later use in reaction based enumeration.

2.6. Mining for Reaction Sequences. For each reaction experiment we recursively searched for previously and subsequently performed experiments. To this end we recursively looked for experiments where the product of the previous reaction acted as reactant. Similarly we searched for experiments where the reactant of the current reaction was product. Here we restricted ourselves to the master reactant, i.e. the reactant that was the basis for yield calculation. In both cases a time limit of 30 days for two subsequent reaction steps was applied for performance reasons. If several identical subsequent or previous reactions were obtained only one representative was taken. The recursion was stopped in case no experiments were found anymore within the 30 days limit. Starting from an experiment a tree of subsequent as well as previous reaction sequences is obtained. Experiments already captured have to be monitored in order not to account several times for certain branches of the tree. The trees of previous and subsequent reaction sequences were then combined in all possible ways into complete reaction sequences, i.e. multistep syntheses.

The following steps were performed separately for all three sizes of reactive centers (*small*, *medium*, and *large* SMIRKS reaction descriptions). For each reaction experiment in a reaction sequence the previously derived SMIRKS (without atom map indices) and the corresponding reaction type group information were looked up. The SMIRKS strings (without atom map indices) of each reaction step were concatenated to yield an identifier for the reaction sequence. Based on this reaction sequence identifier it was monitored on the fly how often a certain sequence occurred. Furthermore, the number of distinct reactants for each reaction step within a specific reaction sequence was monitored in order to calculate a variability score (see eq 1). Postprocessing in KNIME[23] included filtering out reaction sequences with missing SMIRKS information for certain reaction steps or zero overall yield as well as grouping by concatenated consolidated reaction type names for analysis purposes (see section 3.1 and Figure 10).

2.7. Building a Retrosynthesis and Enumeration Tool. Based on the processed reaction and reaction sequence information a tool for retrosynthetic analysis and reaction based enumeration was built. The various steps involved in the tool are depicted schematically in Figure 3, whereas Figure 4 shows screenshots illustrating the user perspective. Starting from a user supplied query molecule the program loops through all available retrosynthetic reaction types (*large* SMIRKS patterns derived from reversed reactions, see above). If no match is found, the program exits. In case of a match all reaction sequences where the matching reaction type appears in are looked up and applied in retrosynthetic fashion to the query molecule. Only if the complete sequence can be applied to the query molecule it will be considered. The resulting retrosynthetic analyses are presented to the user in order of increasing resulting fragment size and decreasing overall yield. The user can then select to see the complete reaction sequence of one or several of the presented retrosynthetic analyses. These may simply correspond to the reversed retrosynthetic analyses; they may however also add further synthetic steps. In the latter case the query molecule appears somewhere in the middle of the complete reaction sequence. The third step consists of choosing one or several reaction sequences and enumerating a library for the selected reaction sequence(s). Starting from the fragments generated from the query molecule by retrosynthesis, each SMIRKS of the reaction sequence is applied to the molecules subsequently in the forward (synthetic) direction. Always one reactant position (in all possible positions) is filled with the starting fragments (in the first step) or the intermediate molecules (subsequent steps). For the other reactant molecules the program looks up the reactants corresponding to the current reaction type (SMIRKS) stored in SMILES format. In this way the generated molecules will have certain features coming from the retrosynthetic fragments of the query molecule as well as they will be diverse due to the other reactants used. The resulting library is written to an SDfile.[25] For each generated compound it contains a descriptive molecule name which allows the user to track the different reactants used, the sequence of consolidated reaction type names, and an example reaction sequence with associated yield to allow look-up of reaction protocols in the ELN, as well as the frequency of the reaction sequence and its length.

3. RESULTS AND DISCUSSION

We describe a procedure that allows one to automatically analyze the knowledge contained in electronic laboratory notebooks (ELN). Central to the procedure is a reduction of all reactions to their reactive center allowing for classification and subsequent reaction type based analyses. Furthermore, knowledge about reaction sequences is extracted and classified. The extracted knowledge is then combined to build a retrosynthesis and reaction based enumeration program for *de novo* design of compounds which have a high chance of being synthetically accessible.

3.1. Analysis. Grouping similar reactions together based on their reactive center led to a considerable data condensation. Figure 5 shows the percentage of reactions covered as a function of the number of reaction types. Here, a reaction type is defined by a certain SMIRKS pattern describing the reactive center (see Figure 2 and section 2.4). Whereas relatively few reaction types are sufficient to describe a large part of all reactions, considerably more reaction types are necessary to cover all reactions in the ELN. Rarely occurring reaction types which contribute little to the coverage can stem from a variety of reactions. These reactions can correspond, e.g., to rarely used, but valid chemistry. However, also an unusual way of inputting a reaction into the ELN may lead to generation of a separate reaction type by the algorithm although the reaction

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

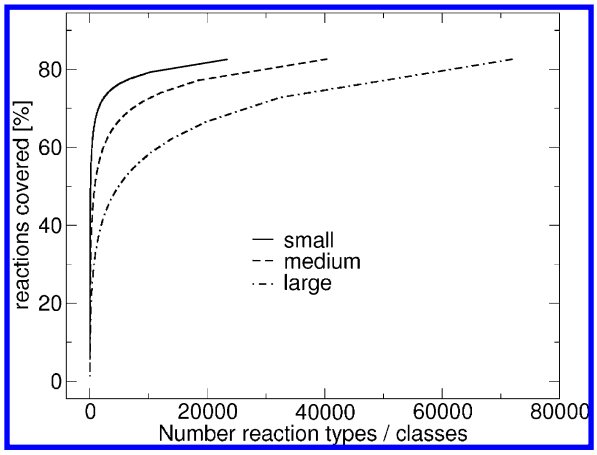


Figure 5. Percentage of reactions covered as a function of number of different reaction types for different reaction center sizes (*small, medium, and large*). A large part of all reactions is covered by relatively few reaction types.

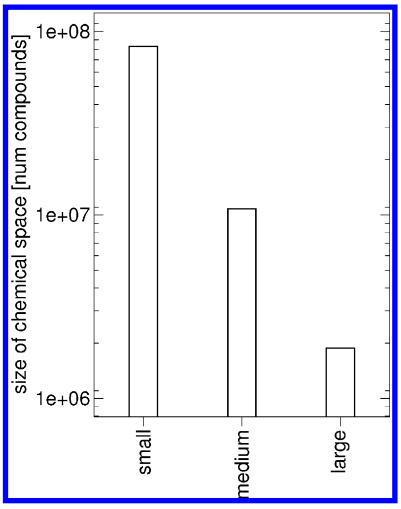


Figure 6. Size of the chemical space estimated as the sum of all possible virtual product molecules of all reaction types obtained. The size of the chemical space decreases with increasing reaction center size (*small, medium, large*). However, the reliability of synthetic accessibility is expected to increase with increasing reaction center size. The space could be further enlarged by adding other than the actually used reactants to the reactant lists. Including reaction sequences would also yield a considerably larger space.

might have been manually classified as belonging to a common type. Depending on the reactive center size all 332 168 reactions for which SMIRKS were derived are described by 23 304 (*small* reactive center), 40 196 (*medium*), and 72 256 (*large*) reaction types, respectively. This corresponds to a data condensation down to 7%, 12%, and 22% which is somewhat less than the data condensation of 6% obtained by Law et al.[20] when deriving reaction rules from the complete Beilstein database. When the analysis was restricted to reaction rules occurring at least five times, Law et al. obtained a condensation down to 2% comparing well to 1.2%, 2.2%, and 3.1% obtained by us for the *small*, *medium*, and *large* reaction center sizes. However, far fewer reaction types are sufficient to cover a substantial part of all reactions. This effect is most pronounced for the least specific (*small*) reaction description where 1500 reaction types cover 70% of all reactions (see Figure 5). This high coverage by only relatively few reaction types is most likely not to be expected for the Beilstein database.

The reaction center specificity has a strong influence on the size of the virtual chemical space spanned by the extracted reaction types and the corresponding reactant molecules. A rough estimate of the size of this space can be obtained by counting the number of distinct reactants one, two, and so on for each reaction type group. Multiplication yields the expected number of virtual products for a certain reaction type. An estimate for the size of the chemical space is then obtained by summation over all expected virtual products of all reaction types. Figure 6 shows this estimate for the different reaction center sizes. The size of the chemical space decreases with increasing reaction center size, whereas the reliability of the synthetic accessibility is expected to increase. Note that we only estimate the size of the virtual chemical space accessible through single step reactions. Including reaction sequences would yield a considerably larger space.

Grouping of reactions based on their reaction center description (SMIRKS) was used to investigate reaction type frequency and group properties. Figure 7 illustrates how often which type of reaction appeared in the investigated data set (using reaction centers of *medium* size). The top 15 reaction types (SMIRKS) are listed in Figure 7 (a). For each reaction type a name is automatically assigned corresponding to the most frequently used name for this type of reaction. The reliability of this consolidated reaction name increases with the frequency of a reaction type. Furthermore, group properties were calculated. We show the average yield in a group as well as a score which judges the potential for diversification (see eq 1). Within the top 15 reaction types, different types of amide bond formation as well as one type of nucleophilic aromatic substitution are scored as diversifying. In order to obtain a more intuitive view of the frequency distribution of various reaction types, Figure 7 (b) is based on a second grouping according to the consolidated reaction type name. In this view it can again be seen that a large part of all reactions is covered by only few reaction types — the top ten cover 41%. Similarly to the findings of Roughley and Jordan[27] in their recent analysis of medicinal chemistry reactions, amide bond formation is predominant covering 13% of all reactions (16% "N-acylation to amide" in ref 27.). The different types of amide bond formation — as defined by their SMIRKS string — are split up in Figure 7 (c). The most common type was found to be the reaction of a secondary cyclic aliphatic amine with an aromatic acid and it is depicted in Figure 7 (d). Some reactions could not be included in the analysis (see Figure 7 (b)). Nine percent of all reactions were discarded as not suited mostly due to missing information on product yield and amount. Another 8% could not be processed mainly due to partial or total lack of atom map indices. Remapping these reactions would be an option to recover a certain part. However, due to the large total number of successfully processed reactions this was not carried out.

Most reactions performed by medicinal chemists consist of several reaction steps. Therefore, we decided to go beyond single reactions and search for reaction sequences in the ELN. Recursively tracking where products were reactants and vice versa should give a good impression on frequency distributions of synthesis length and sequences. However, it should be noted that the reaction sequences obtained in this way need not correspond to a sequence of reactions performed subsequently by one person in one laboratory. Figure 8 shows a frequency distribution of the number of synthetic steps estimated as described in section 2.6. The frequency clearly decreases with

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

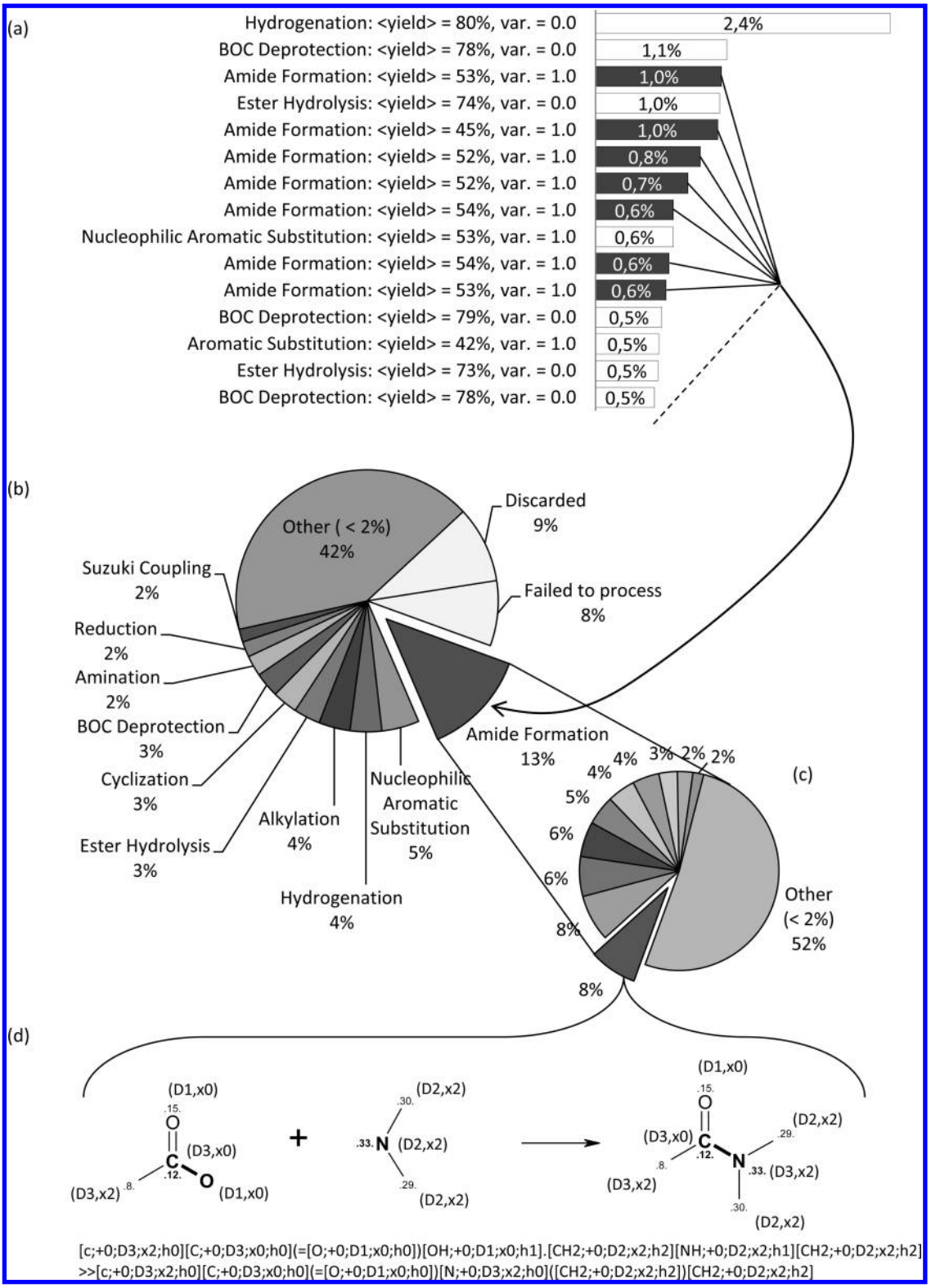


Figure 7. Frequency analysis of different reaction types using reaction centers of *medium size* (see Figure 2). (a) The fifteen most frequent reaction types. Each reaction type is defined by a SMIRKS string describing the reaction center. The reaction name shown corresponds to the most frequently used name for reactions yielding a certain SMIRKS (consolidated reaction type name). Furthermore, the average yield per reaction type and a variability score (see eq 1) are shown. The variability score describes the potential for diversification, ranging from 0 indicating no diversification potential to 1 indicating high diversification potential. Variability scores are rounded to two significant digits. (b) Grouping together all reaction types having the same consolidated reaction type name: Ten reaction types cover a large percentage of all reactions. (c) About half of all amide bond formations are described by ten SMIRKS patterns (reaction types). (d) Depiction and SMIRKS string of the most frequent type of amide bond formation.

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

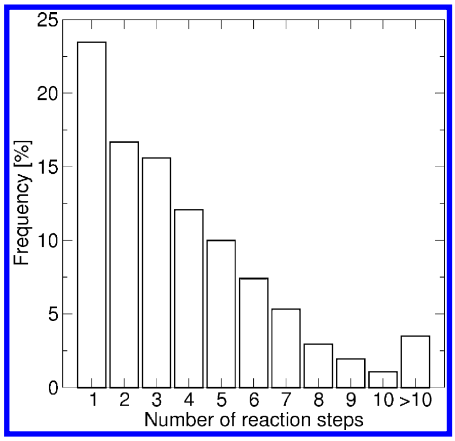


Figure 8. Distribution of number of synthetic steps analyzed as described in section 2.6.

increasing number of steps. The distribution deviates markedly from the one obtained by Roughley and Jordan (See Figure 12a of ref 27.). They find a distribution peaking at three synthetic steps. In the SCRIPDB,[28] which assembles reactions from the patent literature, the distribution of reaction steps shows a decreasing frequency with synthesis length (See Figure 5 of ref 28.). However, the prevalence of single step reactions is much more pronounced than encountered in this study (72% versus 23%).

The coverage of reaction types and corresponding reactions by reaction sequences of different lengths is shown in Figure 9.

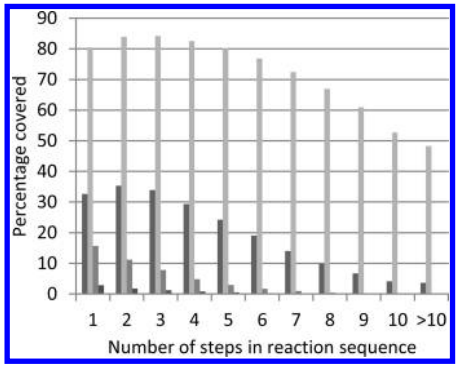


Figure 9. Percentage of reaction (types) covered by reaction types (*medium*) occurring in reaction sequences of different lengths. For each number-of-steps-category the first bar describes the percentage of reaction types covered by reaction types occurring in sequences of that length, whereas the second bar shows the percentage of reactions these types cover. The third and fourth bars show these values for reaction types which only occurred in a single sequence.

For each category the first and second bar show the percentage of reaction types and reactions, respectively, which are covered by the reaction types occurring in reaction sequences of that length. For example, the reaction types (*medium* sized SMIRKS patterns) occurring in single step reactions cover 33% of all reaction types. These 33% percent of all reaction types describe 80% of all 332 168 reactions for which SMIRKS were derived. When looking at singletons, i.e., reaction types occurring only in one sequence, this picture is inverted. For single step reactions, e.g., the singletons cover 16% of all reaction types which, however, only describe 3% of all reactions.

The top five sequences of reaction types for two and three step reactions are shown in Figure 10. For a more intuitive view, the reaction sequences are grouped based on the concatenated consolidated reaction type name. The reaction types involved mainly stem from the top ten reaction types shown in Figure 7 (b). As expected, the percentage of all sequences covered by a certain reaction sequence type is much smaller than for single reactions.

3.2. Retrosynthesis and Reaction Based Enumeration Tool. For maximum reliability concerning synthetic accessibility, the most specific, i.e. *large*, reaction center description (see Figure 2) was chosen for the *in silico* (retro-)synthesis tool. Reaction center descriptions were only considered if the SMIRKS occurred at least five times and was able to regenerate the original product. Successful regeneration was observed for 98% of the SMIRKS fulfilling the frequency criteria. Failures were mainly due to multiple distant changes occurring in one molecule such as a coupling combined with a deprotection. These types of reactions would need to be split up into multiple successive steps in order to describe them by SMIRKS reaction transforms. Requiring SMIRKS to fulfill the criteria in synthetic and retrosynthetic direction yielded 10 110 *large* SMIRKS reaction descriptions. Setting the additional requirement that the SMIRKS must appear in a valid reaction sequence (see section 2.6) reduced this number to 8 393 reaction descriptions (SMIRKS) which cover 66% of all reactions for which a SMIRKS string was derived.

On the basis of these reaction descriptions (SMIRKS) with corresponding reaction sequences as well as the extracted reactant molecules, a retrosynthesis and reaction based enumeration tool was built (see section 2.7). In the following the results of one example application will be presented. The top ranked virtual synthesis of the query compound[29] is shown in Figure 11. The suggested route is an established benzimidazole synthesis. Capital font indicates diversifying steps. ELN identifiers for an example sequence with high yield and product amount are supplied for look-up of experimental procedures. Here, a series of experiments is shown which gave an overall yield of 75%.

Following the user selected synthetic route(s), a library of analogues accessible via equivalent chemistry can be enumerated. Choosing the top ranked route resulted in a library of 29 920 compounds. Figure 12 shows distributions of Lipinski properties as well as similarity to the query molecule for the generated library. Combinatorial enumeration led to molecules with a wide range of properties which are rather dissimilar to the query compound. The properties are predominantly influenced by the reactant molecules extracted from the ELN and to a lesser extent by the query molecule. Here, some compounds are very lipophilic and have rather high molecular weight. However, many molecules also have lower ClogP[11] and a molecular weight below 500 Da. One could filter for simple additive properties prior to enumeration, but as molecule generation is fast enough to run in interactive fashion we suggest post filtering according to desired properties. An application scenario would be the generation of analogues accessible via equivalent chemistry and postgeneration filtering for compounds with an improved predicted property profile. Properties of interest could include predicted affinity, physicochemical properties such as solubility, or ADMET related properties such as, e.g., hERG activity.

Nearly all of the *de novo* generated molecules in the presented example are novel in the sense that they could not be found in compound databases such as PubChem[30] or SciFinder.[31] Figure 13 shows the library core and examples of

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

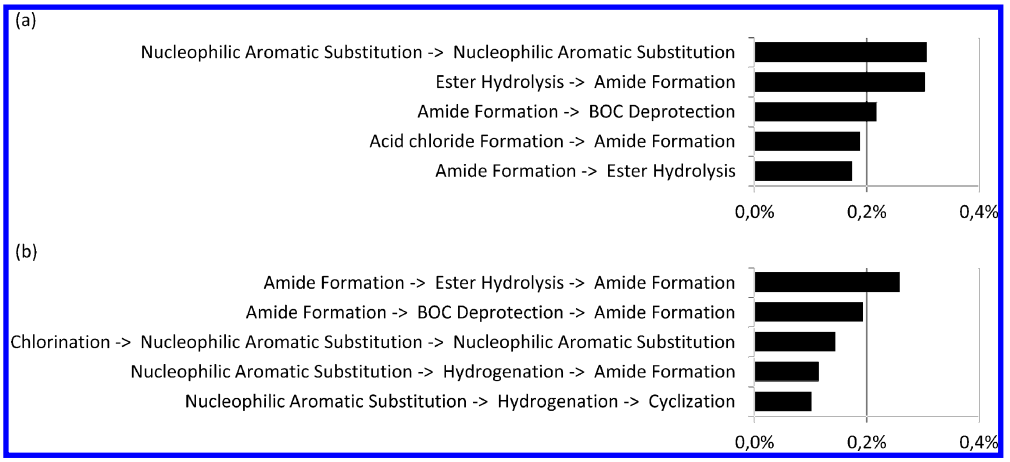


Figure 10. Five most frequent types of two and three step reactions as defined by concatenated consolidated reaction type names (see Figure 7).

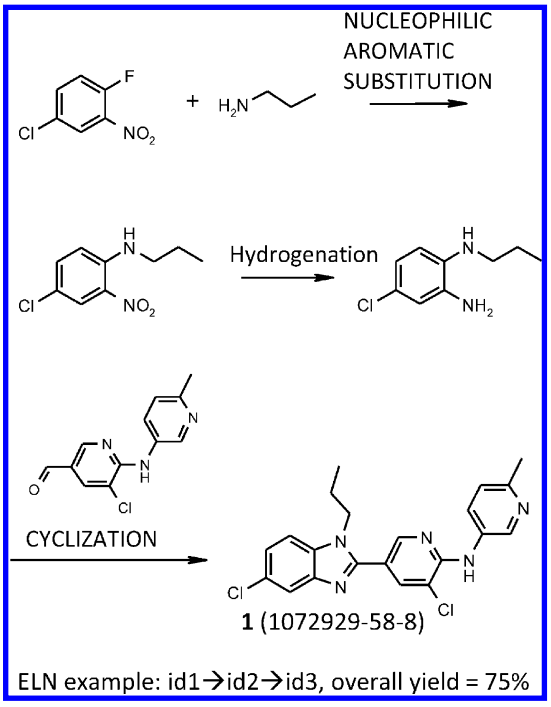


Figure 11. Top ranked synthetic route to compound **1**[29] suggested by the retrosynthesis and reaction based enumeration tool (*large* reaction center sizes, see Figure 2). Compound **1** is not contained in the electronic laboratory notebook from which the knowledge for the tool was derived. Consolidated reaction type names (see Figure 7) are written in capital font if the variability score (see eq 1) is greater than zero indicating the steps in the reaction sequence where diversity is introduced. The user is further supplied with an example reaction sequence from the electronic laboratory notebook (ELN) which allows the look-up of further information (identifiers depicted as id1-id3).

the 23 products which were found in a SciFinder search. A PubChem search did not yield any additional hits but delivered 22 of the 23 compounds found in SciFinder. Hence, 99.9% of the generated molecules are novel in the sense that they are unknown to the public domain.

The retrosyntheses and reaction based enumeration tool presented above is designed to maximize synthetic accessibility of the *de novo* generated compounds. This is not only reflected by the use of *large* reaction center descriptions but also by the choice of reaction sequences and reactants.

We chose to enumerate along observed sequences of reaction types. This is in contrast to iteratively applying a randomly selected reaction of all matching reactions as, e.g., in SYNOPSIS.[9] The latter procedure, which could be easily implemented using the SMIRKS extracted in this work, would sample a greater part of chemical space and could suggest more novel synthesis ideas but possibly at the risk of suggesting very unlikely synthetic routes. An intermediate procedure would be to calculate transition probabilities from one reaction type to another from the knowledge base and use these to select the next reaction to be applied.

Unlike in other approaches[9,12] we decided to apply reaction transforms not directly to the query molecule itself but to a precursor. The tool retrosynthetically goes back to points in the synthesis where variability can be introduced. The subsequent reaction based enumeration allows introducing diversity in a way that mimics very closely the approach of a synthetic chemist in the laboratory.

Another restriction we have chosen for maximum synthetic accessibility was to solely use reactants which have been encountered in the ELN for the employed reaction type. This mitigates the fact that, similar to other approaches,[12] the reaction transform descriptions used (SMIRKS) are local substructure descriptions and therefore do not account for influences of distant functional groups. If arbitrary molecules were permitted as reactants, implementing an estimate of reactivity[19] would be beneficial. SYNOPSIS uses a rule based expert system to judge whether a reaction can be applied. Manually coding reactivity rules is, however, not feasible for the large amount of reaction rules extracted in our approach. An option would be reactivity predictions based on machine learning approaches which, however, suffer from lack of negative data, i.e., reported instances of nonsuccessful reactions.[32]

4. CONCLUSIONS

We have presented an approach to automatically extract and use the knowledge contained in electronic laboratory notebooks. It allows us to answer the questions we put, concerning, e.g., reaction frequency and average yield or the identification of diversifying steps, in multistep synthesis. Furthermore, our procedure gives access to the large chemical space of virtual molecules which are in principle accessible with the chemistry contained in the investigated ELN. In a straightforward procedure reactions are reduced to their reactive center

dx.doi.org/10.1021/ci300116p | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

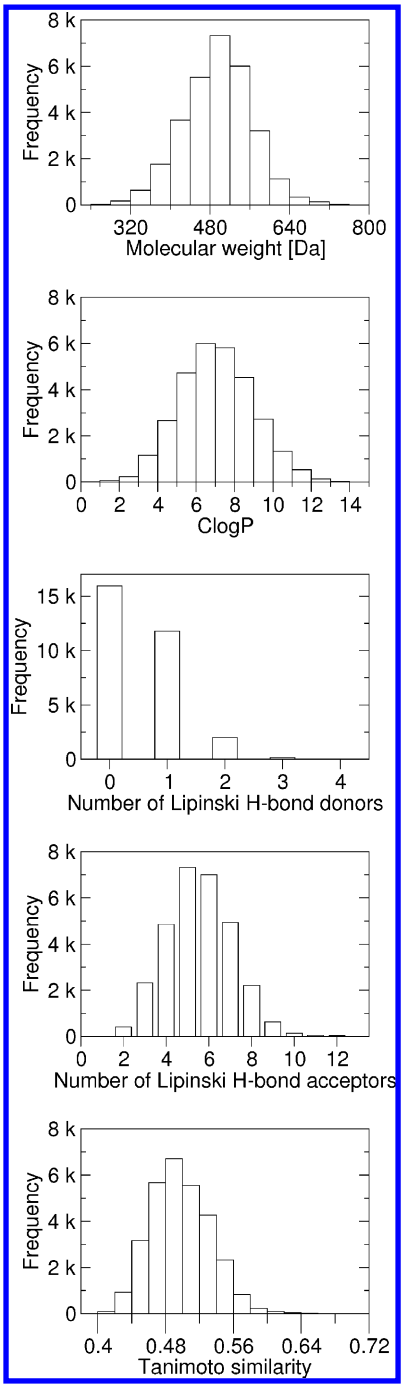


Figure 12. Property distributions of the generated 29 920 molecules: Molecular weight, ClogP,[11] Number of Lipinski H-bond donors and acceptors, and similarity (Tanimoto, Daylight fingerprints[11]) relative to compound **1** (see Figure 11). Note that most of the logP values predicted to be above seven are unrealistic in nature.

allowing for classification and reaction type based analyses. No manual intervention is necessary which makes updates easy and the method applicable to large scale databases of reactions. The extracted reaction rules (SMIRKS) can be used for classification purposes but also to perform reactions *in silico*. If a reaction rule requires additional reactants we use molecules extracted from reactions that resulted in this reaction rule. This limitation of reactant space improves synthetic accessibility of the *de novo* generated compounds as it mitigates the fact that the reaction transforms (SMIRKS) do not account for influences of distant

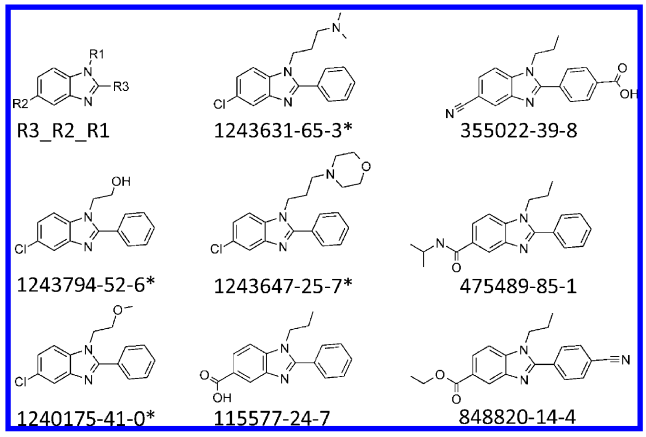


Figure 13. Core describing the library of the 29 920 generated molecules and selected examples of the 23 substances that were found using SciFinder.[31] CAS Registry Numbers are given, and commercially available substances are marked with an asterisk. Compound names returned to the user indicate the different reagents used (R3_R2_R1 e.g. CDDAZU00737_CCKAZU01501←CCKAZU01504←CCKAZU01507←Query_LB2TOM01006 for compound 1243631-65-3).

functional groups. In view of synthetic feasibility, reaction rules are not applied randomly to the query molecule, but observed sequences of reaction rules are followed. Our (retro-)synthesis tool first retrosynthetically fragments the query molecule. Only in a second step reaction rules are applied in synthetic fashion. Starting with a retrosynthesis allows going back to points in the synthesis where diversity can be introduced. Going back to these points enables generation of a diverse set of compounds in a way which mimics closely the approach a synthetic chemist would choose. In an example application the tool generated 29 920 molecules with diverse properties. Almost all of the molecules (99.9%) are novel in the sense that they are unknown to the public domain.

AUTHOR INFORMATION

Corresponding Author

*E-mail: clara.christ@boehringer-ingelheim.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Michael Bieler, Christoph Hoenke, Cyrille Kuhn, Alexander Weber, and Nils Weskamp for helpful discussions. Valuable feedback on the manuscript from Herbert Köppen and Domnic Martyres is gratefully acknowledged. Furthermore, we would like to thank our colleagues in compound management and information systems for technical support.

DEDICATION

C. D. Christ would like to dedicate this work to her inspiring and supportive mentor of many years, Wilfred F. van Gunsteren, on the occasion of his 65th birthday.

ABBREVIATIONS USED

ELN, electronic laboratory notebook; CASD, computer-aided synthetic design

REFERENCES

(1) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(2) Kutchukian, P.; Shakhnovich, E. De novo design: Balancing novelty and confined chemical space. *Expert Opin. Drug Discovery* **2010**, *5*, 789−812.

(3) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.* **2011**, *672*, 299−323.

(4) Boda, K.; Johnson, A. Molecular complexity analysis of de novo designed ligands. *J. Med. Chem.* **2006**, *49*, 5869−5879.

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

(6) 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.

(7) Podolyan, Y.; Walters, M.; Karypis, G. Assessing synthetic accessibility of chemical compounds using machine learning methods. *J. Chem. Inf. Model.* **2010**, *50*, 979−991.

(8) Kutchukian, P.; Lou, D.; Shakhnovich, E. FOG: Fragment optimized growth algorithm for the de novo generation of molecule: Occupying druglike chemical Space. *J. Chem. Inf. Model.* **2009**, *49*, 1630−1642.

(9) Vinkers, H.; De Jonge, M.; Daeyaert, F.; Heeres, J.; Koymans, L.; Van Lenthe, J.; Lewi, P.; Timmerman, H.; Van Aken, K.; Janssen, P. SYNOPSIS: SYNthesize and OPti mize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765−2773.

(10) 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.

(11) *Daylight Theory Manual*, 2011. www.daylight.com/dayhtml/doc/theory/index.html (accessed Nov 4, 2011).

(12) Patel, H.; Bodkin, M.; Chen, B.; Gillet, V. Knowledge-based approach to de Novo design using reaction vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163−1184.

(13) Hristozov, D.; Bodkin, M.; Chen, B.; Patel, H.; Gillet, V. J. Validation of Reaction Vectors for de Novo Design. In *ACS Symposium Series*; American Chemical Society: 2011; Vol. *1076*, pp 29−43.

(14) Broughton, H. B.; Hunt, P. A.; MacKey, M. D. 2003, US 2003/0182094 A1.

(15) 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.

(16) Cramer, R.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and searching 10^{20} synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 341−350.

(17) Boehm, M.; Wu, T.-Y.; Haussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.* **2008**, *51*, 2468−2480.

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

(19) Todd, M. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247−266.

(20) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Knew, S.; Johnson, A.; Major, S.; Wade, R.; Ando, H. Route designer: A retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **2009**, *49*, 593−602.

(21) *OEChem*, version 1.7.2.4; OpenEye Scientific Software Inc.: Santa Fe, NM, 2009.

(22) *Oracle C++ Call Interface*. www.oracle.com/technetwork/database/features/ oci/index-090820.html (accessed Nov 4, 2011).

(23) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Studies in Classification, Data Analysis, and Knowledge Organization; Springer: Berlin, Heidelberg, 2008; pp 319−326.

(24) *Pipeline Pilot*, version 8.0; Accelrys: San Diego, CA, 2010.

(25) Dalby, A.; Nourse, J.; Douglas Hounshell, W.; Gushurst, A.; Grier, D.; Leland, B.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at molecular design limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(26) *ICClassify*. http://infochem.de/content/downloads/classify.pdf (accessed May 4, 2012).

(27) Roughley, S.; Jordan, A. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54*, 3451−3479.

(28) Heifets, A.; Jurisica, I. SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents. *Nucleic Acids Res.* **2011**, *40*, D428−D433.

(29) Carcache, D.; Vranesic, I.; Blanz, J.; Desrayaud, S.; Fendt, M.; Glatthar, R. Benzimidazoles as potent and orally active mGlu5 receptor antagonists with an improved pk profile. *ACS Med. Chem. Lett.* **2011**, *2*, 58−62.

(30) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; American Chemical Society: Washington, DC, 2008; Vol. *4*, pp 217−241.

(31) *SciFinder*, version 2011; Chemical Abstracts Service: Columbus, OH, 2011.

(32) Carrera, G.; Gupta, S.; Aires-de Sousa, J. Machine learning of chemical reactivity from databases of organic reactions. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 419−429.