appears as: Customizable Generation of Synthetically Accessible, Local Chemical Subspaces, Joshua Pottel and Nicolas Moitessier
Journal of Chemical Information and Modeling 2017 57 (3), 454-467
DOI: 10.1021/acs.jcim.6b00648

Exploring Synthetically Accessible Chemical Space.

Joshua Pottel[‡] *and Nicolas Moitessier**

Department of Chemistry, McGill University, 801 Sherbrooke Street W., Montréal, Québec, Canada H3A 0B8.

nicolas.moitessier@mcgill.ca

Abstract

Screening large libraries of chemicals has been an efficient strategy to discover bioactive compounds, however a portion of the potential for success is limited to the available libraries. Synergizing combinatorial and computational chemistries emerged as a time-efficient strategy to explore the chemical space more widely. Ideally, streamlining the evaluation process for larger, feasible chemical libraries would become commonplace. Thus, combinatorial tools and, for example, docking methods would be integrated to identify novel bioactive entities. The idea is simple in nature, but much more complex in practice; combinatorial chemistry is more than the coupling of chemicals into products: synthetic feasibility includes chemoselectivity, stereoselectivity, protecting group chemistry and chemical availability which must all be considered for combinatorial library design. In addition, intuitive interfaces and simple user manipulation is key for optimal use of such tools by organic chemists and for the integration of such software in medicinal chemistry laboratories. We present herein FINDERS and REACT2D—integrated into the VIRTUAL CHEMIST platform, a modular software suite. This approach enhances virtual combinatorial chemistry by identifying available chemicals compatible with a user-defined chemical transformation and by carrying out the reaction leading to libraries of realistic, synthetically accessible chemicals—all with a completely automated, black-box, and efficient design.

INTRODUCTION

After years of development, docking-based and ligand-based virtual screening techniques are now commonplace in medicinal chemistry.[1-4] Currently, large molecule databases enable efficient screening of millions of purchasable compounds.[5] One might consider, however, that these are relatively limited virtual libraries, restraining computational approaches from reaching their true potential: screening hundreds of millions, or billions, of synthetically accessible compounds as shown by Hartenfeller and co-workers.[6] In practice, synthetically accessible chemical space is attainable, but cannot be realistically experimentally explored nor physically stored on a shelf. Using computational methods to guide combinatorial chemistry was popular in the late 90's and early 2000's, focusing on maximizing the quality of the combinatorial chemistry using chemoinformatics.[7] However less has been devised to generate synthetically accessible libraries of chemicals, including virtual *de novo* library design techniques.[8] These approaches connect the chemical space accessible in the wet lab to the screening/storage space available with computers. For example, a team of researchers at Pfizer,[9] Hartenfeller *et al.* from Novartis[6, 10] and Masek and co-workers at Certara[11] have focused on the virtual design of synthetically accessible libraries by applying chemical transformations commonly used in medicinal chemistry. Specifically, the Pfizer Global Virtual Library (PGVL) has a number of features: reagent compatibility for a given transformation, the removal of protecting groups on the final product, the consideration of 3-component reactions largely used at Pfizer, and a very large database of reactions.[9] However, this package, developed over more than a decade, is not available to the medicinal chemistry community.

Ideally, a broadly applicable tool to access unexplored virtual chemical space could be developed for all chemists, not just experts in computational methods. The situation can be likened to NMR spectroscopy; its integration into the organic chemist's toolbox has significantly improved the throughput of routine experiments, while NMR spectroscopists continue to develop advanced methods. Some major disconnects to overcome between computational and organic chemistry are the advanced 3D graphical interfaces and the use of command-line input—both of which often remain unintuitive for non-experts. For example, we previously developed a program, REACT,[12] that generates combinatorial libraries from chemicals drawn in 3D and complex atom-type based chemical transformations. Although this tool was diverse in its applicability, it required technical expertise to encode the chemical transformation in the form of atom type changes. Ultimately, the need for extensive training must be reduced; in contrast, learning to run routine ^{1}H NMR spectroscopy experiments requires only a few hours. In this context, we considered developing simple, intuitive, and automated tools to generate libraries from 2D chemical drawing applications.

Herein we present our efforts that led to the creation of FINDERS, a program for chemical search, and REACT2D, a program for combinatorial library generation. These two programs were then fully integrated into our VIRTUAL CHEMIST platform for asymmetric catalyst design and our FORECASTER platform for drug discovery – both freely available to the academic community.

METHODS

Simulating real experiments. Ongoing research in our laboratories illustrated the appeal of intelligent searching and combinatorial chemistry tools (**Figure 1**).[13] At this stage, we had a novel synthetic method to prepare chiral oxazepanes that could be further developed as asymmetric catalysts or enzyme inhibitors. A common approach would be to use computational methods—docking small molecules for selecting potential enzyme inhibitors. However, drawing all of the synthetically accessible analogues would be tedious and prone to errors. Moreover, querying a database of purchasable compounds would yield few hits due to the novelty of the chemistry. Ideally, software could derive the corresponding virtual combinatorial library from catalogues of purchasable starting materials, using the chemical transformation, as in **Figure 1**b. To successfully accomplish the task, software should be able to: encode a chemical transformation, be aware of chemical (in)compatibility, and search catalogs for suitable chemicals. Available programs may require the user to provide information on the chemical centers (labeling atoms of reactants and products), but chemical transformations as drawn by chemists do not readily include this information. We proposed to develop a tool that can "understand" the chemical transformation for a scheme typically drawn by organic chemists without needing extraneous information.

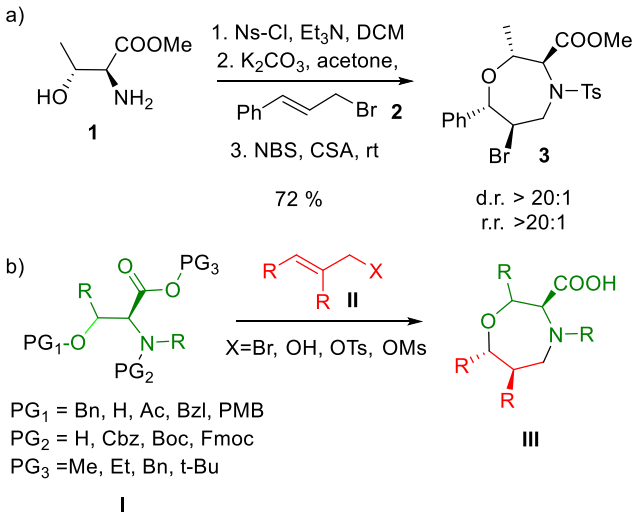


Figure 1. a) Developed chemical synthesis of 1,4 oxazepanes to be used as organocatalysts or enzyme inhibitors. b) A generalized synthetic scheme to be interpreted by software for accessing uncharted areas

of chemical space.

Encoding and manipulating chemical structures

The problem. Computationally searching and transforming chemical structures in an automated fashion is best summarized by managing substructures—both common and identical; threonine methyl ester (**1**, Figure 1) is a chemical that matches the general structure **I** (reagent search) while **I** (green) is incorporated into the final product **III** (green).

Molecular representation. Generally, from an algorithmic perspective, finding a substructure within a structure is known as sub-graph isomorphism. A graph, in computer science, is very similar to a molecule in the sense that it has vertices (atoms) and edges (bonds) and, all together, these describe the graph (molecule). There are several types of graph-matching: exact (a complete match), substructure (a partial match) and similarity (minor differences). The three endeavors each have an intrinsic complexity due to the encoding of molecules.

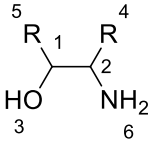
An organic chemist handles chemical structures in 2D, as shown in Figure 1 and usually wants to find most, if not all, matches of a chemical scaffold with other variable features, referred to as R groups. The information can be stored on a computer as a text file of atom names, coordinates and connections in various formats (e.g., Structure Data Format (SDF)).[14] Some further encoding of this representation is required to establish comparisons useful for substructure matching. SMILES[15] is one method of keeping all the molecular information in one text string, something useful for exact matching, however there can be inconsistencies from one output source to another, as two different strings can code for the same molecule. The IUOAC International Chemical Identifier (InChi)[16] is another representation which is unique for any compound. Further technical difficulties arise when adding the ambiguities and variability for substructure and similarity searching. As a result, we thought to develop a minimalistic approach to encode molecular structures into a string-like format that could then be compared from one to another using a breadth-first search (BFS) algorithm. This problem is known to be NP-Complete[17] meaning no polynomial-time solution is known. Pragmatically, this suggests that simple heuristics will be required to shorten the runtime. Fortunately, much of the sub-graph isomorphism problem applied to chemistry is made simpler by the constraints imposed by the nature of molecules.[18] Certain rules exist that, for example, enforce a limit on edges that touch a given vertex which allows us to take shortcuts and greatly reduce the running time.

Thus, all that is required as input should be a 2-dimensional drawing of the molecule, stored as text in one of many formats.[19, 20] The atoms, their numbers and elements, as well as the bonds, their numbers, their types and the atoms they join, can all be interpreted directly. This simplistic approach is very naïve however its crude approximations enable rapid comparisons of one structure to another. The string representation (genotype) then encodes the necessary information from the actual molecule (phenotype). We planned to only perform manipulations/computations based on this genotypic representation as opposed to accessing data structures—phenotypes—during each computation. The text-based encoding would then be used to identify scaffold matches, exact matches and largest common substructures between molecules, processes necessary to search for chemicals and encode chemical transformations.

Filtering molecules. Since we aim to manage millions of molecules, our software requires intelligent pre-processing in order to exclude as many irrelevant comparisons as possible. Many of the shortcuts to filtering non-matches and the attempts to identify exactitudes as efficiently as possible are based on the approaches of the human brain. A chemist can look at two molecular drawings and almost immediately label them as identical or not. A computer should replicate these efficiencies. First, before computationally considering the specific properties of the genotypic representations, conceptually, a chemist would look for easily identified, superficial features.[21] If well-selected, this should reduce the runtime of the software significantly. The key notion is to label independent properties that will cover the widest possible range and that will be computationally easy to compare. For example, if the number of atoms and bonds differ from one molecule to the next, then they cannot be exactly the same. Similarly, if they do not contain the same number of each element then they are not identical. Terminal

atoms—those that are bound to only one other atom—are also an easy target to distinguish dissimilar molecules. Other properties such as patterns, rings and other functional groups (sets of atoms)[22] have been shown to function well however we deem them too costly to compute with our atomistic approach. The advantage to filtering in this manner is that only the template must remain in memory; each test molecule is read one at a time, processed and then written or discarded. Any molecules that pass the filter are output to a file and not stored in memory during the execution of the program. We believe that the space requirement could be astronomical if every structure passes the filter and thus some speed is sacrificed in return. Later, these structures will again be read in, one at a time, and again only one molecule will be stored at a time.

If the first filters are passed, the genotypic string is created to represent the molecular structure. When considering two images, one technique to eliminate possible similarity is to start in a distinct region that is likely to differ from one to the other. In this context, the string is rooted, or anchored, at one of the rarest elements appearing in the template—that which is the least abundant in the molecule—and is then expanded in a breadth-first-search manner (Figure 2). In the second layer, the neighbors of each atom in layer 1, except those appearing in the previous layer (layer 0) will be appended, and so on to eventually create the complete graph/molecule. Within the string, only specific information for defining any atom is kept: the parent, the element, the atom number, the degree of un-saturation and a flag for cycles. While more information may be required to describe more complex systems, our current testing does not demonstrate a need for any additions. With this genotype in hand, structures can now be compared with various goals: substructure, exactitude and largest common substructure identification.



a) anchor atom:

Element	Atom number	Unsaturation	Ring bridge	Layer end
N	6	0	0	:

b) following atoms

Parent	Element	Atom number	Unsaturation	Ring bridge	Layer end
6	C	2	0	0	:

c) complete string ',' if in same layer

N;6;0;0:6;C;2;0;0:2;R;4;0;0,2;C;1;0;0:1;R;5;0;0:1;O;3;0;0:

Figure 2. Example of indexing an amino-alcohol into its genotypic representation.

Substructure search. Finding a substructure (e.g., template with R groups) within a structure (e.g., actual chemical from a library) is crucial for chemical search. In this context, R-, Ar- and G-groups are defined for the variability of the search; they are common placeholders for: any combination of atoms starting with a carbon, aromatic groups, and *any* combination of atoms respectively. While the definition of R varies from place to place (R may stand for any functional groups, or be more restrictive with R being only carbon chains), organic chemists traditionally use R to describe aliphatic chains (and sometimes including R=H) and Ar (Aryl) for aromatic groups. In the case of R=H, some explicit rules were required, as the string representation described above does not include hydrogens. We have added G as generic groups. For example, usual functional groups on an aromatic ring—such as a nitro group, a methoxy group and a methyl group—are connected to the ring through C-C bonds or bonds with atoms other than carbons. These would be designated as G. In contrast, the first atom of an R group in R-CH_2-X used in S_N2 reactions is most likely a carbon atom.

The library phenotypes must be converted, one at a time, to their genotypes as was demonstrated above for the template. As the string is built, at each layer, the two strings will be compared (Figure 3). If, in this comparison, discrepancies arise, the following atom is tried as the anchor and a new string is

built (each molecule has various genotypes, similar to SMILES strings). If potential matches are not discovered after querying each atom in the substructure (after examining all possible genotypic representations) then this library molecule does not contain the template. If the library string is completed and all template atoms have a potential match, it progresses to the next stage of evaluation. To clarify, it cannot yet be concluded that the library molecule contains the template since the strings are built in a forward manner – meaning information of the past is lost; each atom "knows" only to whom they are connected and nothing beyond (Figure 3). There are consequently instances that require a more rigorous investigation.

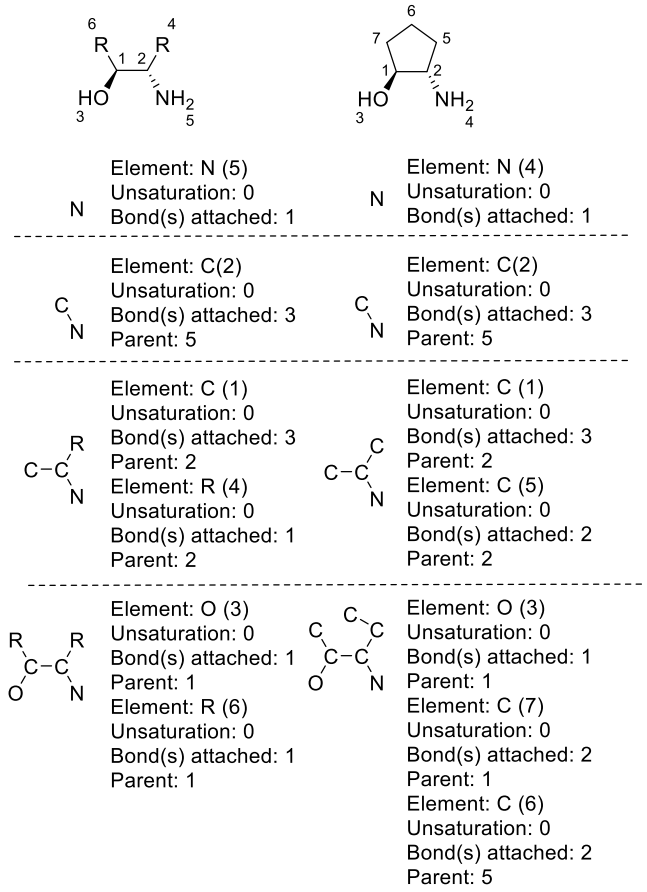


Figure 3. A sample run-through of the BFS matching algorithm (only 3 iterations). Each atom in the template (left) has a potential match in the query molecule (right).

A series of deductive tests are carried out to establish the unique corresponding atoms between template and library molecules. Each template atom must have at least one unique counterpart in the library molecule to ensure correctness as well as to properly label the equivalencies. This is achieved by creating new string genotypes anchored at atoms in the template that have more than one potential equivalent atom in the library molecule. If the new genotype does not match the corresponding genotype created from the library molecule, anchored at the potential matching atom, then these atoms cannot be counterparts. In essence, this approach results in the molecules being evaluated forwards *and* backwards. If, after any iteration, one template atom is no longer matched to any library atom, the entire molecule does not match and is skipped. A proper substructure is identified if and only if each and every template atom is labeled with at least one unique counterpart. This approach accounts for ring junctions (special treatment for 3-member rings) as well as local symmetries found within many molecular structure.

Exact structure matching. All of the components presented in the previous section are pertinent to exact structure matching except no R groups are present and every atom in the template must match

only one in the library molecule. It is in fact much simpler algorithmically to determine if two molecular structures are identical. No consideration of tautomers was included in the development at this stage.

Largest common substructure identification. Establishing the largest common substructure found between two molecules (Figure 4) was a challenging task. Unlike the previous two search goals, an anchor is not obvious, the number of corresponding atoms in both structures is variable, and the criteria for equivalency is less stringent.

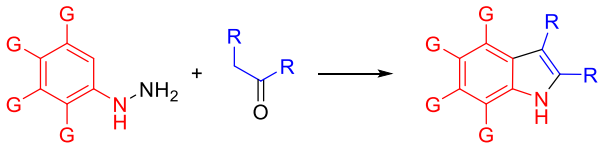


Figure 4. Largest common substructures identified for two reagents within a product. Atoms or bonds in black are not part of the largest common substructure.

In Figure 4, the G and R groups can be established as the anchors for the two reagents respectively. Upon building and comparing genotypes, many leniencies are allowed when searching for equivalencies. The atoms connected to a given atom, the level of unsaturation of bonds and the ring junctions can all differ between two molecular substructures yet they can still be common substructures. To circumvent these variables, the genotype construction is slightly modified to consider number of bonds rather than unsaturations; if this and the element match, it is preliminarily set as a potential counterpart. Using the above-mentioned extensive graph matching anchored at "promiscuous" atoms (with multiple equivalencies), the result is narrowed to the maximum number of atoms that are labeled with one single counterpart. The entire largest substructure procedure is repeated from each potential anchor (G and R atoms) in order to achieve the largest possible match.

IMPLEMENTATION

Searching for chemical scaffolds. FINDERS (Filtering, Identifying, Negating Duplicates and Evaluating Reaction Substructures) is an automated sequence of algorithms that has been implemented to search through user-defined catalogs based on an input reaction scheme; an organic chemist can draw a 2D scheme in the sketcher (Figure 5), A+B→C (A→B, A+B→C+D and A→B+C also implemented), and run the software package in order to extract libraries of reagents matching scaffolds A and B—all in a matter of seconds or minutes depending on the generality of the scaffolds and size of the catalog. It is important to note that the rigorous structure matching is what differentiates this approach from internal searches of other databases—some of which ignore one, or several of stereochemistry, aromaticity, symmetry, and generalized groups—usually to avoid sacrificing speed.

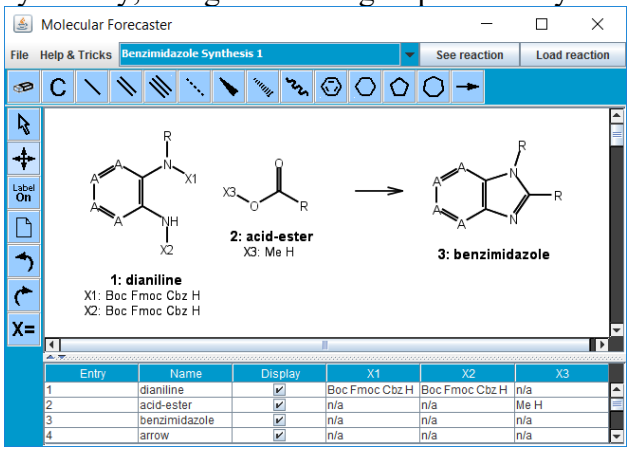


Figure 5. Benzimidazole synthesis scheme drawn with the implemented sketcher.

Within the oxazepane synthesis in Figure 1, amino alcohols are required; however, diversity can be expanded if protected amino alcohols are also considered. Practically, this is interesting to an organic chemist due to variations in cost, availability, and ease-of-use. For example, chemists in our lab would consider Boc-protected amino alcohol with diverse R groups if the unprotected analog were not available. A single, often high yielding step would make the unprotected compound synthetically available. Algorithmically, the given scaffold is expanded for protecting/leaving groups (Figure 6) by the developed algorithm: CREATE (Chemical Reagent Expansion After Template Evaluation). Up to 4 "X-groups" are allowed and a list of available options is given in **Error! Reference source not found.**. These groups need to be explicitly defined since they differ from R, G and Ar-groups. Each generated scaffold is now queried amongst the filtered libraries for substructure matches using the theory mentioned above and output to a new file.

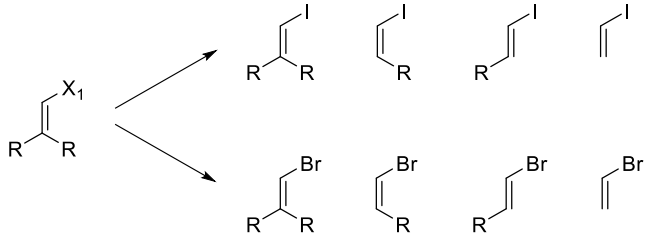


Figure 6. An example of CREATE expanding templates to include leaving groups, iodine and bromine in this instance, and R being a functional group or H.

Organic chemists are selecting chemicals based not only on similarity, but also on functional group compatibility. For example, in a reductive amination reaction ($R^1CHO + R^2NH_2 \rightarrow R^1\text{-}CH_2\text{-}NH\text{-}R^2$), the generic group of the aldehyde (R^1) should not include: an amine (would react under reductive amination conditions), another aldehyde (polyreaction may occur), an alcohol (may form a lactol which may reduce the reactivity of the aldehyde), an acyl choride (would react with the amine to form an amide) or several other functional groups. Simply enumerating the chemicals would therefore lead to unrealistic chemistry. To address this issue, we have also integrated other in-house programs to label functional groups (SMART – Small Molecule Atom typing and Rotatable Torsions assignment), and remove user-defined incompatible chemical groups on the generic chains (REDUCE – Recognition and Elimination by Descriptors of Undesired Chemical Entities) and added the necessary menus in our interface (Figure 7). In order to further reduce the library to potentially drug-sized molecules, chemicals can also be filtered by molecular weight or Lipinski's rule of five.

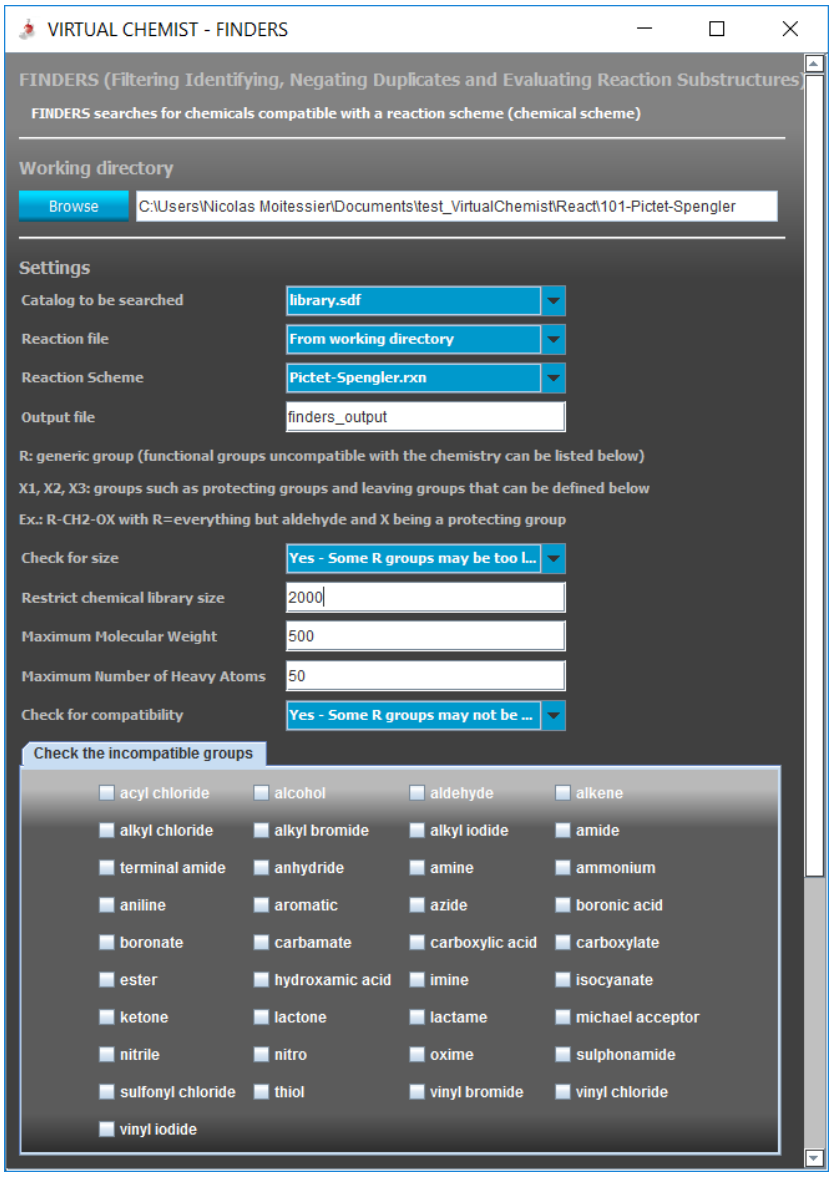


Figure 7. The visual interface allowing users to restrict chemical size, library size and identify incompatible functional groups from a list of options.

At this stage, a library of chemicals compatible with the chemical scheme is generated. However, for the chemistry shown in Figure 1, serine in various quantities (e.g., 1 g, 5 g package), forms (neutral, hydrochloride salt, solution) were identified—molecular duplicates. The next step is therefore to remove these duplicate structures; keeping a single copy. The developed algorithm DIVERSE (Duplicate Identification Validated by Evaluation of Regio- and Stereochemical Exactitudes), uses the exact structure matching to evaluate all remaining molecules and eliminate duplicated structures. This is often the most time-consuming step since an exponential number of molecular comparisons are required in its current state. To reduce this impact on runtime, each molecule of a catalog is translated to a bit string describing its formula and presence of unsaturations and subsequently, only those with identical bit strings are compared. It is worth mentioning that some software packages, such as eMolecules,[23] use pre-processed molecules to further reduce the search time as well.

Finally, SELECT,[12] a program using MACCS keys for clustering and extracting the most diverse chemicals, is applied (optionally). This ensures that if, for instance, an A+B→C reaction is used, huge libraries of products are avoided which may cause issues with storage/runtime.

Performing combinatorial chemistry. REACT2D (Rapid Enumeration by an Automated Combinatorial Tool in 2D) uses two chemical libraries to carry out an A+B→C type of reaction and generate all the possible combinations of A and B forming C (or carry out an A→B reaction, generating all products B, A→B+C generating all B and C, and A+B→C+D generating all C and D) again, in a matter of minutes. It was designed to be an independent set of algorithms since properly focused libraries may already be in-hand (and not coming from FINDERS). For this reason, many of the algorithms are re-run with some minor modifications. First, CREATE expands the scaffolds once more, but in this instance, leaving/protecting groups are labeled as X after being matched since these groups will not appear in the final product C and would have the same combinatorial outcome. Thus, when DIVERSE is carried out, the same scaffolds differing only by a leaving/protecting group would be labeled as identical so as to not unnecessarily grow the number of combinations. Furthermore, it is applied at this stage and not after the combinatorial chemistry since the number of molecular comparisons is far fewer prior to combinatorial enumeration. While some of the current available software requires the user to explicitly label atoms belonging to the different reactants in the product, our implementation precludes human interference.

To summarize, the substructure matching has, again, identified the corresponding atoms between template and library molecules, the exact matching has removed ones that would result in duplicate products and now the reaction can be carried out. Using the theory on largest common substructure, the reacting templates, A and B, can be matched to the atoms found in C. The leniencies defined regarding largest common substructure are due to bonds forming/breaking and rings being closed from reactant to product. Finally, using matrix algebra, library molecules can be templated onto the products C and be joined in a combinatorial fashion.

As discussed below, some care must also be taken to ensure that the 2D geometry of each product generated is realistic. Algorithms have been implemented to ensure that angles, bonds and rings geometry is correct and stereochemistry is maintained.

Stereochemistry, aromaticity and symmetry. While other programs have been described,[9, 10] little attention has been put on stereochemistry and aromaticity—major issues to be considered. It is well known that enantiomers or diastereomers of a drug may have significantly different bioactivity and/or toxicity among other properties and should be distinguished. Additionally, in catalysis, the stereochemistry of an organocatalyst is vital to its purpose.

While stereogenic centers may be formed during the course of the reaction, some stereogenic centers may also be present in the reagents. Upon combination, the orientation of the reagents may change and the integrity of the stereochemistry must be maintained. For example, as illustrated in the amide bond formation in Figure 8, if the reagents are properly oriented, the solid wedge bond should be constant during the course of the reaction (Figure 8a). However, if some orientation changes must be carried out (Figure 8b), modification of the symbol (solid wedge to hash) may be necessary to maintain the stereochemistry. Geometrical routines were implemented to consider the rotation and flips of chemicals.

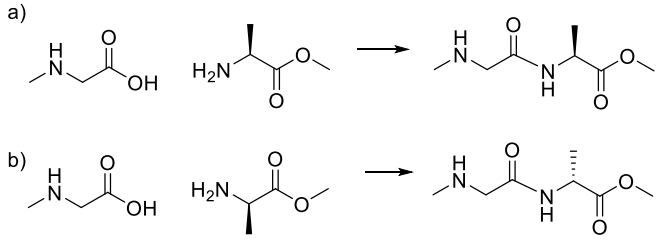


Figure 8. Stereochemistry integrity maintained in two orientations: a) The wedge bond is unchanged due the consistent drawing from reagent to product. B) The wedge bond is changed to a hash bond since the reagent, as drawn, is flipped when overlaid onto the product template.

An additional stereochemistry issue is that some chemicals, even in commercial databases, may be either missing or poorly drawn. In Figure 9, while the left structure is correctly drawn, the other two are

not and must be identified. Once more, routines have been implemented to search for stereogenic centers and whether they have been assigned. In addition to stereogenic carbon atoms, double bonds are stereogenic and their E/Z configurations should be distinguished as shown later with the Stille coupling reaction.

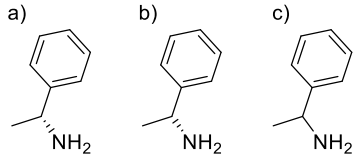


Figure 9. Stereochemistry integrity lacking within databases: a) molecule correctly drawn, b) improper stereochemistry and c) lacking stereochemistry.

Aromaticity is also another factor to consider. Aromatic chemistry is quite different from alkene chemistry. For example, addition of Br_2 to double bond does not occur on aromatic rings (Figure 10a). When searching for alkenes, butene should be identified as a candidate and toluene should not (Figure 10b). In addition when filtering out R groups for compatibility, R should not include another alkene as polybromination may occur. So 2,4-hexene should be filtered out while *trans*-β-methylstyrene should be selected (Figure 10c). Heteroaromatic rings such as thiophene or the isomeric triazoles should also be labeled as aromatic. Within FINDERS and REACT2D, six membered rings with alternating double and single bonds as well as five membered rings with two double bonds and at least one hetereoatom (among O, S and N) are considered aromatic.

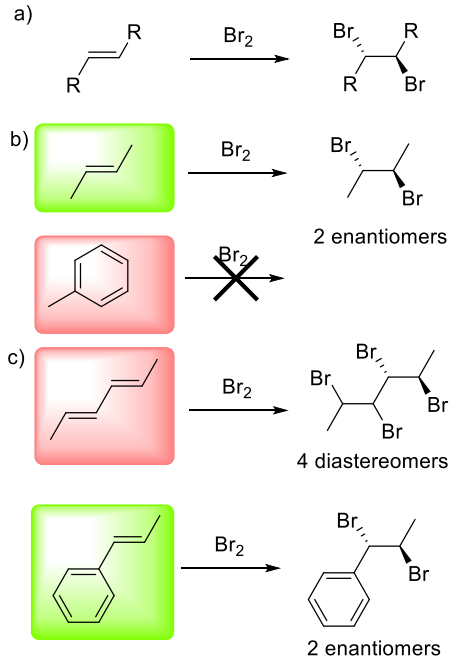


Figure 10. Differentiation between aromaticity and other unsaturations. a) Bromination of a trans-alkene. b) Bromination occurs with butene and not with toluene; the latter should be excluded when searching a library for this reaction type. c) 2,4-hexene would undergo polybromination and should therefore be excluded, but *trans*-β-methylstyrene has only one reacting double bond and should be included.

Symmetry is of significant importance when automating combinatorial chemistry. For example, let us consider the synthesis of benzodiazole shown in Figure 11. The template of reagent #1 can be matched with the monoprotected dianiline shown in two ways. In the case shown in Figure 11b, the Boc group is considered as a protecting group and is removed—the corresponding free dianiline is reacted. Alternatively, the Boc group is matched with R, is no longer considered a protecting group and the

chemical is reacted with no change, leading to a Boc-containing product (Figure 11c). In organic synthesis, amide/carbamates and amines have very different reactivity and should be distinguished. To address this issue, a routine has been implemented that labels nitrogen as either "amine" or "amide" (for amides, carbamates and carbonates). In the example shown in Figure 11, a dianiline with two reactive "amine" nitrogens is desired. If the Boc group is not considered a protecting group and is matched to the R group, this match includes a nitrogen labeled as "amide" and a nitrogen labeled as "amine" and will be discarded. If the Boc is a protecting group, both nitrogens are labeled as amines and the match is retained.

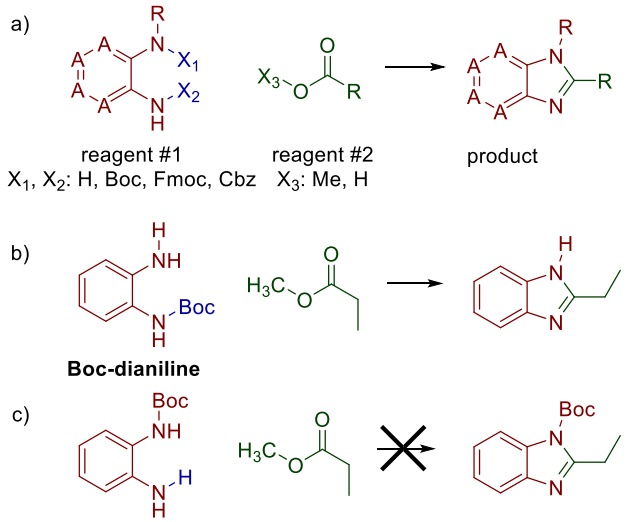


Figure 11. Consequences of symmetry. a) The largest common substructures of two reagents overlaid on the product in a benzodiazole synthesis. b) A candidate reagent where a Boc group is labeled as a leaving group and the reaction is carried out. c) The same reagent is matched differently where R=Boc, leading to a different product; however this reaction would not occur in reality.

Finally, regioisomers should also be produced. In the example shown in Figure 12, the reagent can be matched in two different ways in this A→B reaction. Generally, the symmetry routine ensures that a single match is selected regardless of the different atom numbering and orientation (two versions of the same chemical were drawn). However, there are specific cases (Figure 12b) the two should be kept leading to the two regioisomeric products. The program can be instructed to generate two products (Figure 12a) and will generate both regioisomers

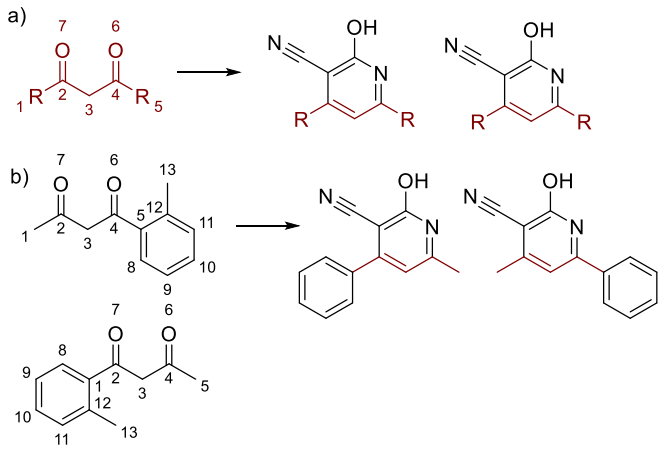


Figure 12. Symmetry in 3-nitrile pyridine synthesis cannot be neglected since two different products can be yielded in specific instances. a) Reaction template with non-obvious relevant symmetry. b) A reagent that would yield two regioisomers after reaction.

Runtime. Molecular substructure search is constrained by the rules of chemistry however these principles also induce complexities beyond those of just vertices and edges. Establishing canonical labeling is a challenging, but necessary venture[24] and doing so requires more information than is superficially provided. In essence, each atom is not only defined by its adjoining atoms, but by their adjoining atoms and effectively the entire molecule. Here, we define every atom only by its element and immediate surroundings. In a linear molecule, like a linear graph, the isomorphism problem is made simpler however those are often not the desired substructures. There have been several difficulties in our approach and we have attempted to address them as they arise in order to never encode more information than that which is required—while never compromising the accuracy of the search. The filtering stage is straightforward however the selection and efficiency of filters depends on the testing set. Certain sets may have more terminal atoms and fewer cycles while other may be exclusively cycles and have many rare atoms such as halogens. In order to optimize the process, statistics could be pre-calculated on a testing set in order to determine which descriptors would be the most diverse and successful at weeding out non-matches. The properties selected were deemed sufficient for this initial version. The running time of the filtering is $O(P \cdot (a + b))$, where P is the number of molecules in the test library, a and b are the numbers of atoms and bonds respectively; essentially it is based on the number of atoms and bonds in each test molecule, i.e. the time to read the information.

The indexing and evaluation stages are, together, more costly in terms of runtime. The worst-case is $O(Q \cdot (a_t^3 \cdot a))$, where Q is the number of molecules in the filtered test library, a_t and a are the numbers of atoms in the test and template molecules respectively. In the worst case, every test structure is explored in full, and all of its atoms are within the same number of layers as the template molecule. This is, however, not often the case and this section of the program tends to run faster than the filtering stage. It is believed that $Q << P$ and furthermore, the a_t term is misleading. Due to the BFS nature of the search (**Error! Reference source not found.**), each test molecule which fails, tends to fail prior to the end of the search and thus the entirety of both a_t and a are rarely fully attained, and in fact, a is only fully explored when a match occurs.

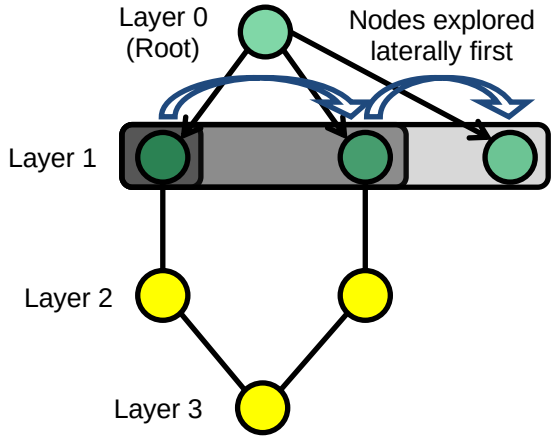


Figure 13. Example of breadth-first search exploring a molecular graph laterally and iteratively rather than the entire molecule immediately.

RESULTS AND DISCUSSION

Chemical search. Searching has evolved over the past century, from file cabinets, inventories and library catalogues to digital databases and online directories, and these changes have required new searching techniques to be developed. More specifically, in the chemical world, the digitization of molecular libraries has made searching more difficult due to the complex nature of chemical structures and information storage. Often, a chemist will want to find most, if not all, matches of a chemical scaffold with other variable features. This desire for chemical diversity spans multiple fields such as medicinal,[25] combinatorial,[26] catalytic[27] and materials[28] chemistries and remains an active field of

research. Tools for chemical database search[29, 30] and virtual combinatorial chemistry have been reported,[9-11] but they are either not available, missing key features such as aromaticity, stereochemistry and/or chemical compatibility, or are not described in details that would allow us to understand their advantages and limitations.

Protocol. Ultimately, two programs (FINDERS and REACT2D) were combined into one automated protocol (Figure 14). First, filters were first implemented into FINDERS. These include simple filters for the presence of simple features such as elements and unsaturations, then filters to remove any chemicals incompatible with the chemistry. A matching algorithm described in the method section is selecting chemicals that match the generic chemicals in the chemical transformation scheme. Duplicates are subsequently removed and the N most diverse chemicals are selected.

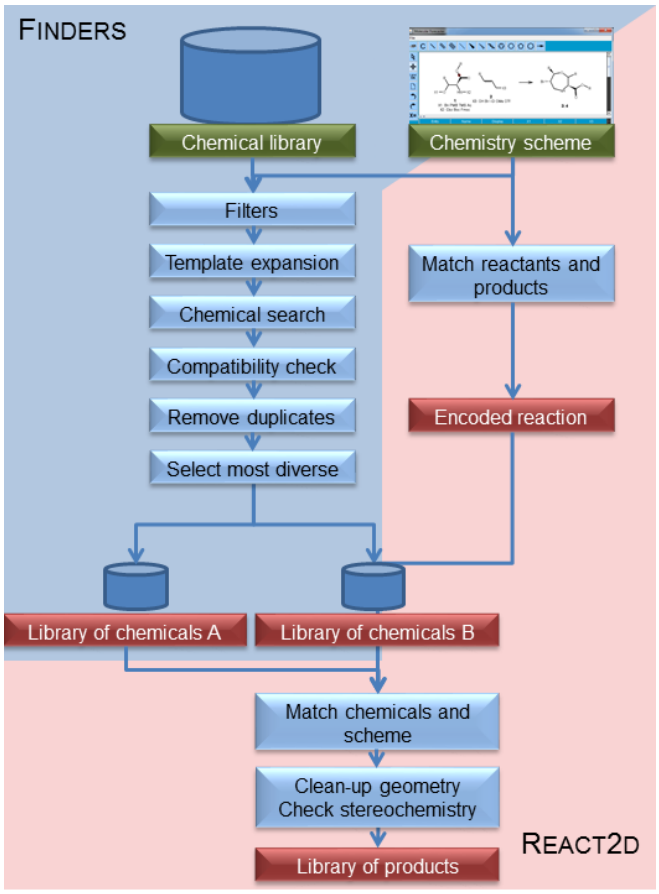


Figure 14. FINDERS and REACT2D automated protocol with intermediate steps labeled.

Validation. A set of twenty-two representative reactions selected from the set of Hartenfeller *et al.*[10] complemented with a few other reactions was used to validate the programs (Table S1). Libraries of 20 chemicals were manually built, one for each of these 20 reactions (available as supporting information). The purpose of a hand-made test set was to select specific chemicals for each reaction that may be difficult for our software to distinguish from one another—each reaction having specific chemistry. The manual construction also ensured certainty that we captured all of the hits and omitted all of the misses.

In order to illustrate the use of FINDERS, consider the Stille coupling reaction (Figure 15). The set of 20 chemicals used to test the program is given. For this specific transformation, we wished to consider the reaction of an alkene stannane, which can be prepared in one step from the corresponding halide derivatives. As a reacting partner, we proposed to use an aryl halide, an aryl iodide or an aryl triflate, the latter being prepared from the corresponding phenol derivatives. In this case, the assignment of aromaticity and double bond stereochemistry was critical. FINDERS was able to identify the 5 chemicals

that matched reagent #1 (green in Figure 15) and the 7 chemicals that matched reagent #2 (red in **Figure 15**). The other chemicals were not selected; either the stereochemistry of the double bond was wrong, the stereochemistry was not defined, or the molecule did not match at all. Out of those selected chemicals, some were filtered out afterwards for compatibility reasons, as required (**2**, **19** and **20**).

With the 4 chemicals as reagent #1 and 5 as reagent #2, REACT2D next prepared the corresponding library of 15 compounds, identifying **10** and **11** as duplicates in the context of this reaction. This example demonstrates some of the strengths of our implementations.

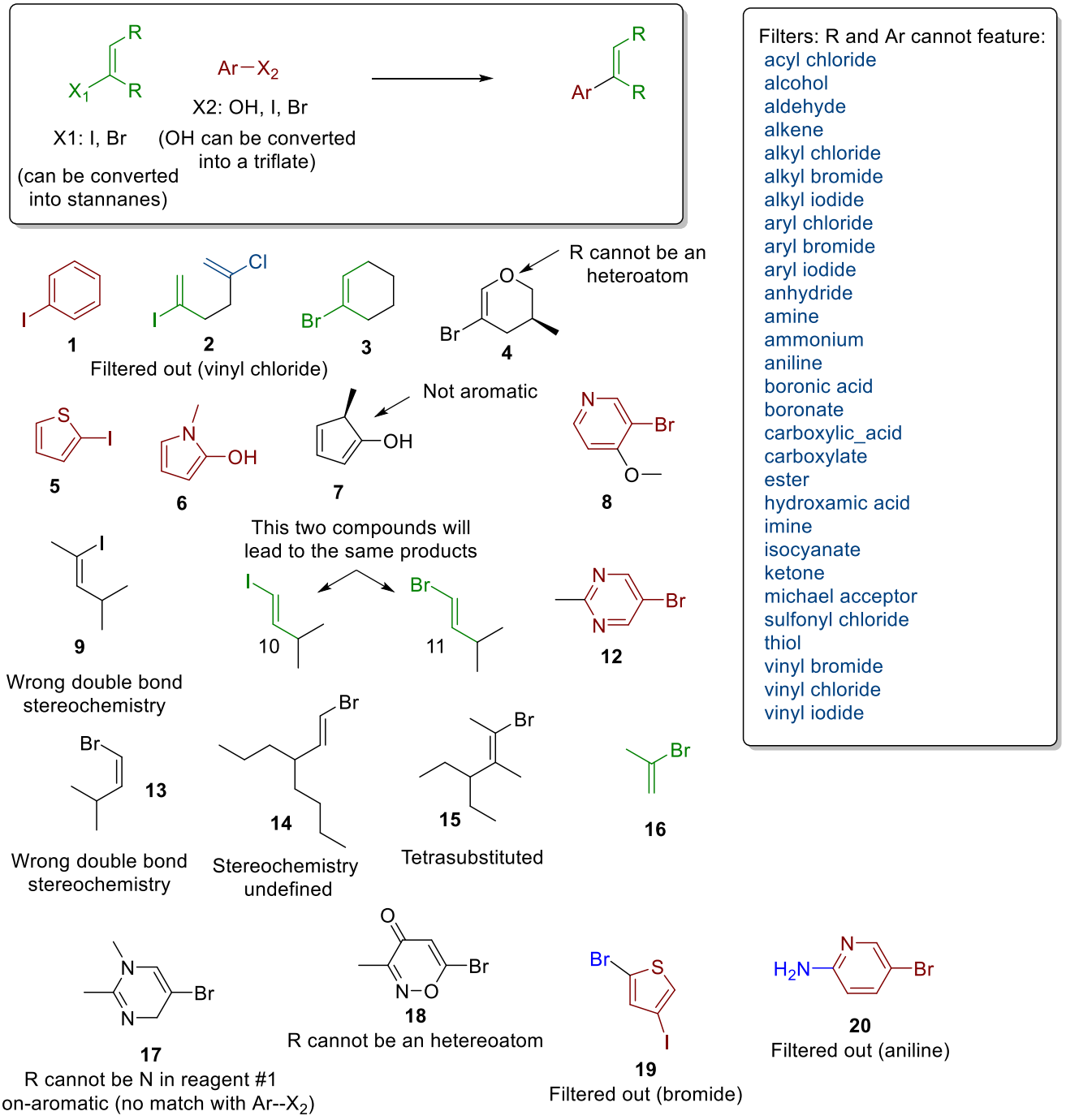


Figure 15. Aromaticity definition and double bond stereochemistry relevant in Stille coupling. Green molecules match the template of reagent 1, red molecules match the template of reagent 2. Explanations for omission labeled under colorless molecules.

Among the stereochemistry issues we had to address are the ones shown in Figure 16. When a stereocenter is formed during the course of the reaction given as template (Figure 16a), this stereochemistry is assigned according to reagents. In the case shown in Figure 16b, a chiral reagent is used and its stereochemistry is retained in the product. The bond carrying the stereochemical information of the reagent (purple sphere) is used to carry the stereochemistry of the forming bond as well (blue sphere). Initial versions of REACT2D were overwriting the reagent stereochemistry with the product stereochemistry (Figure 16b). To address this limitation, the drawing should be adjusted to ensure than the information for both stereogenic centers is retained (Figure 16c). Similarly, the stereochemistry given for an asymmetric reductive amination (Figure 17) reaction applies only to amines reacting with dissymmetric ketones. When an aldehyde or a symmetric ketone is used, no stereogenic centers are formed and both products are equivalent and one should be discarded (Figure 17).

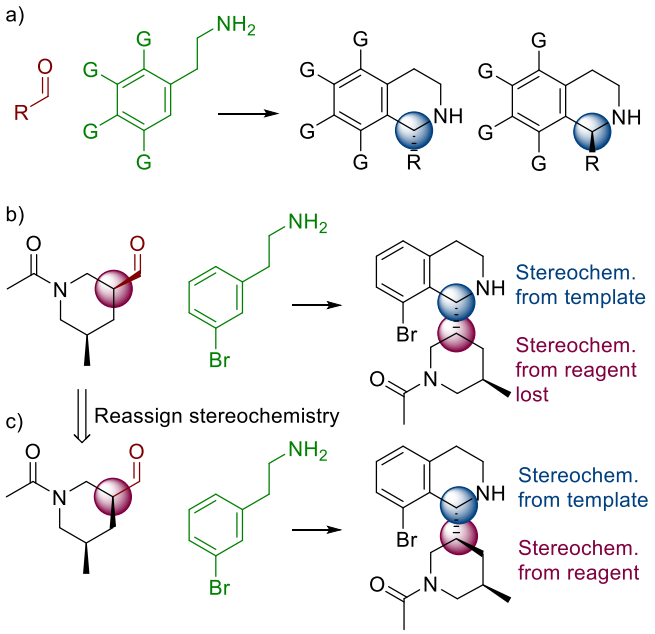


Figure 16. Stereochemistry reassignment for proper drawing. a) Pictet-Spengler reaction template, identifying the generated stereogenic center. b) Stereochemistry in the reagent can be lost c) unless handled correctly.

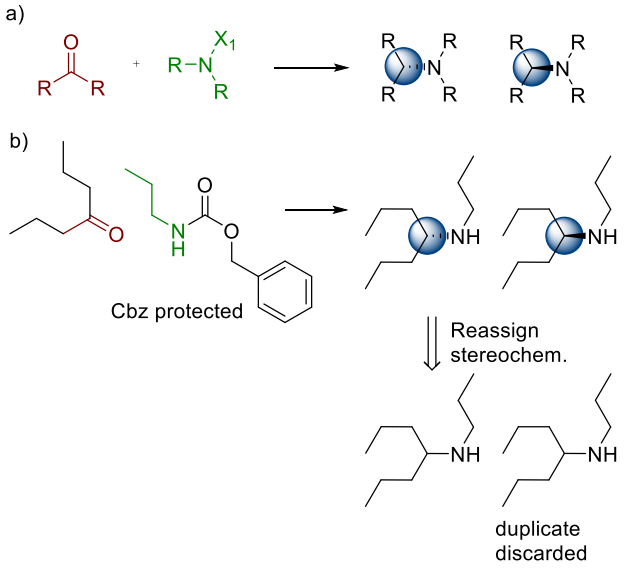


Figure 17. Stereochemistry reassignment for proper drawing and removal of duplicates. a) Reductive amination reaction scheme with generated stereocenter. b) Symmetric ketone leads to non-chiral center

and duplicate products.

As another example of stereochemistry issues, let us consider the Mitsunobu reaction shown in Figure 18a, which proceeds with inversion of configuration. Some of the chemicals shown in Figure 18b should first be identified as duplicates and the chemical library reduced to 8 chemicals (Figure 18c). No stereochemistry should be assigned to dicyclohexylmethanol (**21**) and 2-propanol (**22** and **23**) which are achiral. If one is given, it is ignored. If stereochemistry must be assigned and none is provided as with 2-butanol (**24**), the chemical is discarded. In contrast, if stereochemistry is assigned to a chiral compound, the stereochemistry of the reacting center will be inverted in the course of the reaction as defined in the chemical transformation scheme.

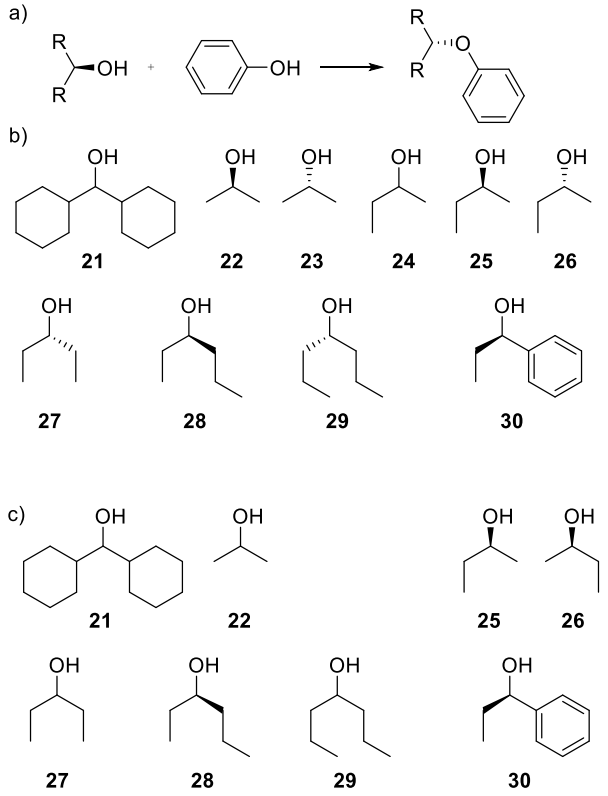


Figure 18. Stereochemistry reassignment for proper matching. a) Mitsunobu reaction. b) library of alcohols; c) stereogenic centers are identified and duplicate compounds together with compounds lacking stereochemical information are discarded.

With these features implemented, when reductive amination—a very common reaction in medicinal chemistry—was selected, the correct library of 42 unique products was built (Figure 19).

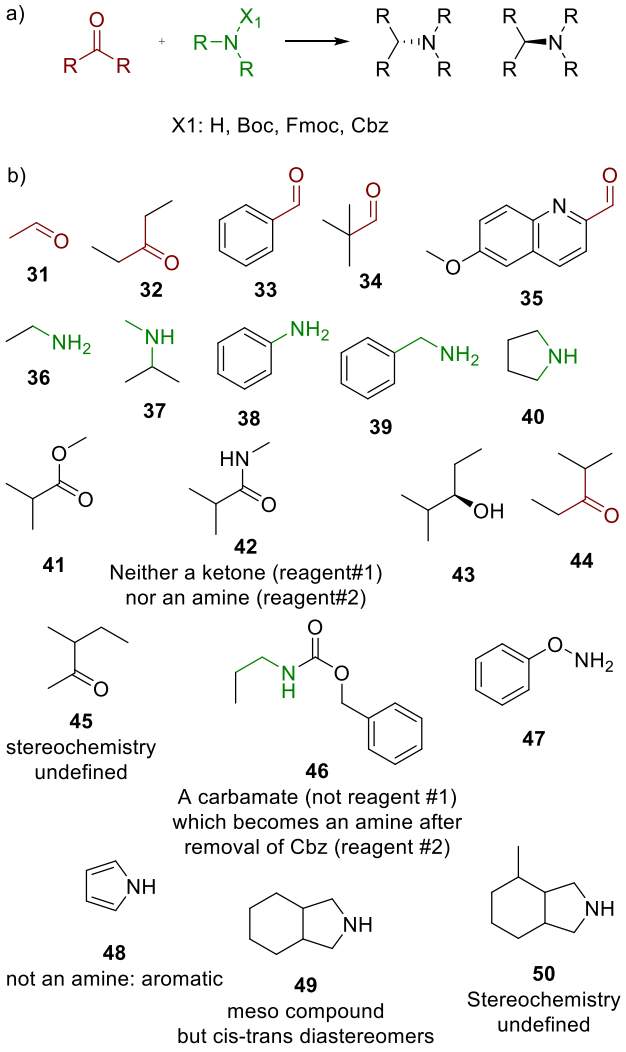


Figure 19. Aromaticity, stereochemistry and "amide" labels relevant in reductive amination. a) Reaction scheme template. b) Library molecules with red and green substructures matching reagent 1 and 2 respectively from the scheme.

Limitations.

Tautomers are difficult to handle with this approach, but should also be considered. For example, the Mitsunobu reaction shown in Figure 20a: if the tautomer 1 is drawn in the synthetic schemes, only the reagents with this tautomeric form will be selected by FINDERS. In addition, if the product 1 is the one given to REACT2D, product 2 will not be generated. As a workaround, Hartenfeller *et al.*[10] proposed to simply use four schemes with the different combinations (searching for reagent with tautomeric form 1 or 2 leading alternatively to product 1 or 2). This alternative approach is reasonable, but considering tautomers in a single run would render the method more user-friendly and robust—similar to the way we have handled stereochemistry and symmetry. However, identifying tautomeric forms of a given chemical is quite a challenge as several functional groups can undergo tautomerisation from simple diketones to heterocycles with concomitant epimerization of stereogenic centers (Figure 21). Within REACT2D, using the two A+B→C+D reaction schemes shown in Figure 20b would suffice.

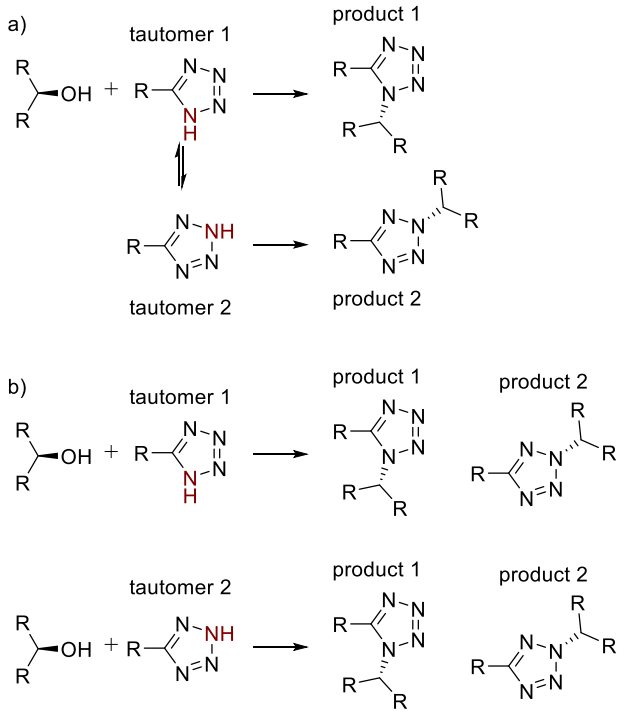


Figure 20. Mitsunobu reaction. a) Multiple tautomers yielding multiple products. b) Current approach in REACT2D without having implemented an automated recognition of tautomerisation.

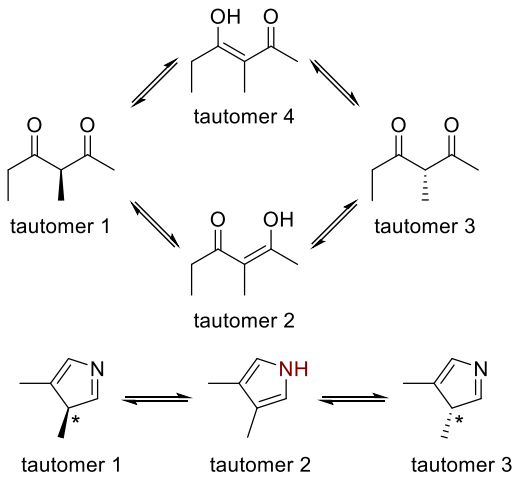


Figure 21. Challenging combinations of tautomers and stereochemistry with no clear resolution.

As another limitation, REACT2D assumes that the reaction undergoes as long as the reagents match the templates. As an example, when the Paal-Knorr reaction is used, FINDERS and REACT2D search for 1,4-diketones. However, when 1,4-cyclohexadione is used, the reaction cannot proceed for geometrical reasons (Figure 22). Although this does not represent a large fraction of the potential reagents, this should be kept in mind if this molecule is later identified as a potential hit.

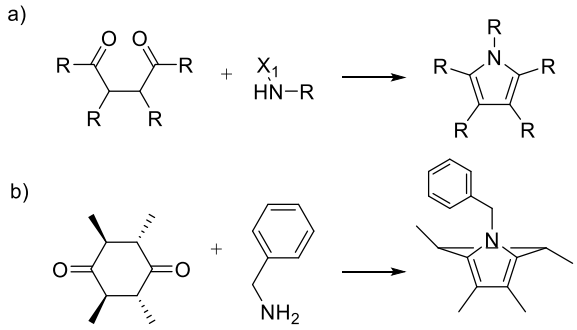


Figure 22. a) Paal-Knorr reaction scheme template. b) A geometric impossibility that would not be captured by the automated protocol.

Finally, some of the stereochemistry elements are significantly more difficult to identify. The quaternary center of the dialkylcyclohexanone shown in Figure 23b is not a stereogenic center. However, when incorporating it into a spirochromanone, this quaternary center becomes a stereogenic center and two stereoisomers can be drawn. This axial stereochemistry is not implemented and the two stereoisomers are considered duplicates.

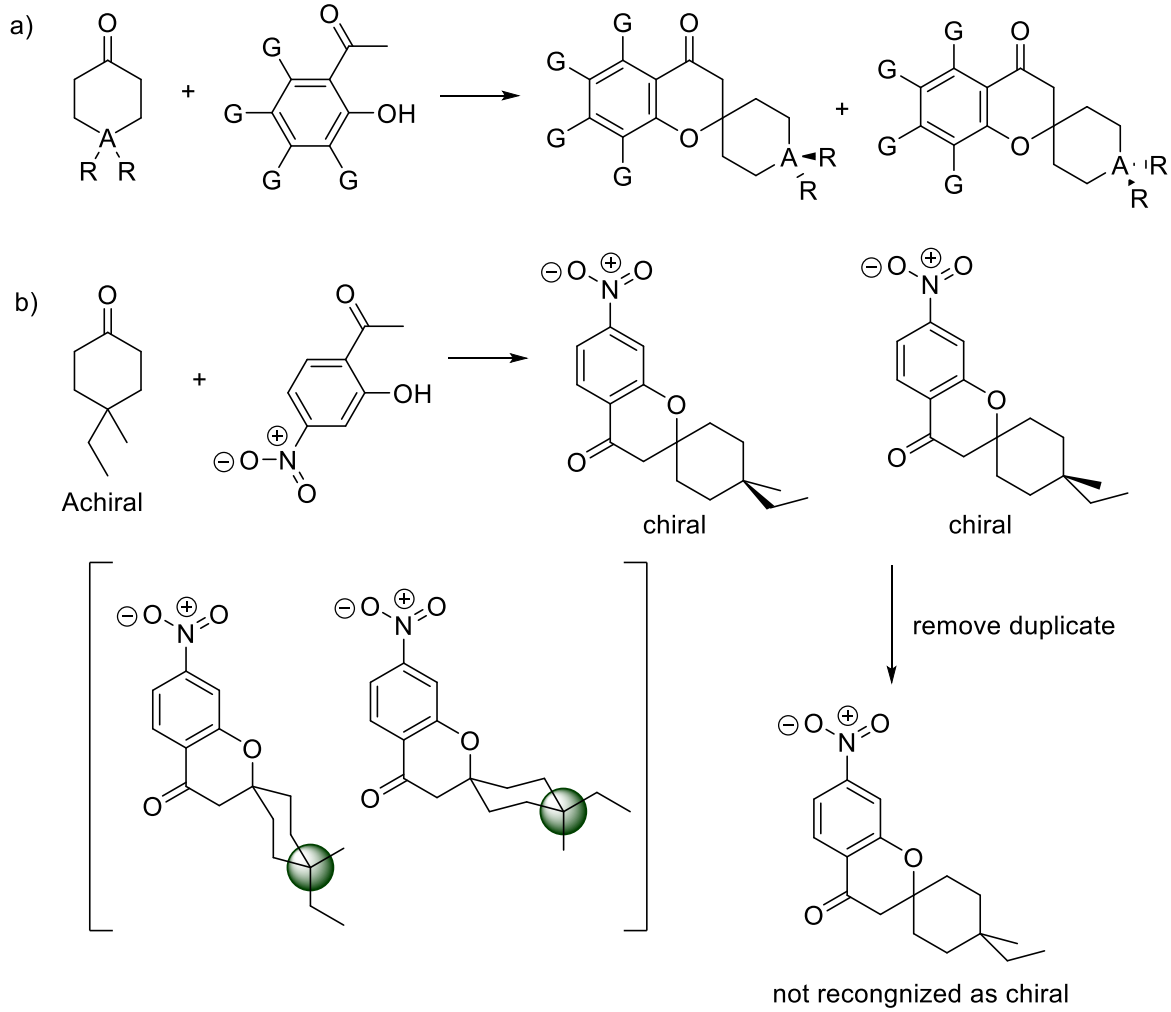


Figure 23. Spirochromanone synthesis and stereocenter generation after reaction. Axial chirality ignored by current version of REACT2D.

Access to chemical space. In order to further assess the power of this software, we used a catalog of ca. 100,000 commercially available chemicals and applied the reactions set from Hartenfeller. In order to limit the library to drug-sized molecules, the chemicals with molecular weight greater than 500 were discarded and the chemical libraries restricted to the 2000 most diverse. The collected data is shown in Table 1.

Table 1. Selected examples for the generation of chemical libraries from a set of reactions and a catalog of chemicals. Data for each of the reactions is given in Tables S4 and S5. The number of molecules remaining after each stage is given as is the runtime.

	FINDERS									REACT2D	
Reaction	After matching		After compatibility		After removing duplicates		After optimizing diversity		Time (min)	Product library size	Time (min)
	A	B	A	B	A	B	A	B		C (+D)	
Pictet-Spengler reaction	3,058	136	2,288	94	1,991	58	1,991	58	7.5	230,935	25.5
Benzimidazole synthesis	260	10,317	219	6,342	91	5,189	91	2,000	47.9	180,000	24.8
Thiazole synthesis	138	113	110	100	98	91	98	91	5.2	8,099	0.4
Nimentowski quinazoline synthesis	45	3,672	43	1,704	30	1,368	30	1,368	12.2	36,936	10.2
Tetrazole synthesis	2,134	-	1,283	-	1,190	-	1,190	-	3.4	2,380	0.8
Triazole synthesis	2,134	172	1,279	136	1,186	129	1,186	129	4.2	2,356,582	191.4
3-Nitrile-pyridine synthesis	233	-	180	-	138	-	138	-	1.3	276	0.0
Spirochromanone synthesis	134	67	119	45	52	40	52	40	2.2	2,080	0.5
Paal-Knorr pyrrole synthesis	12	9,213	11	1,340	5	1,042	5	1,042	13.9	2,072	2.3
Fischer indole synthesis	98	2,625	92	1,734	54	1,055	54	1,055	6.0	56,970	6.8
Oxadiazole synthesis	2,134	8,719	1,281	5,501	1,177	4,545	1,177	2,000	34.5	2,354,000	192.3
Reductive amination	10,620	24,326	3,676	8,040	1,745	3,732	1,745	2,000	122.6	4,796,181	349.2
Suzuki coupling	1,966	11,628	1,245	2,989	1,163	2,757	1,163	2,000	28.9	1,711,647	132.4
Mitsunobu reaction 2	11,944	0	6,460	1	1,166	1	1,166	1	34.0	468	9.6
Stille coupling	78	10,019	20	2,759	17	2,286	17	2,000	29.0	29,546	10.7

Grignard reaction 1	2,134	14,187	841	1,216	767	1,040	767	1,040	19.9	659,620	53.3
Grignard reaction 2	10,620	14,187	5,057	1,203	2,615	1,027	2,000	1,027	50.9	3,295,130	227.2
Schotten-Baumann coupling reaction	11,863	24,326	5,985	12,365	4,879	7,365	2,000	2,000	213.8	4,143,696	314.9
Sulfonamide synthesis	719	24,326	609	12,365	569	7,365	569	2,000	171.9	1,181,813	100.1
Buchwald-Hartwig	24,326	19,145	10,093	5,853	4,942	2,643	2,000	2,000	202.5	4,424,502	530.1
Imidazopyridine synthesis 1	2,509	133	1,356	108	388	95	388	95	34.1	33,626	8.9
Chan-Lam	24,326	1,161	10,093	701	4,942	661	2,000	661	128.9	1,378,740	104.7

FINDERS generated the necessary chemicals for REACT2D for all of the 56 reactions within 43 hours, running serially on a single processor. REACT2D proceeded in 65 hours under the same conditions to generate nearly 40M compounds, which are all synthetically accessible in one step (or two if a protecting group has to be removed). The maximum amount of time for a single reaction was approximately 12 hours for the Buchwald-Hartwig reaction, which yields over 4M products. The majority of reactions require less than 2 hours to complete. These storable libraries could now be converted to 3D and used in docking screens or further filtered for specific properties to generate smaller libraries of potential catalysts.

Interestingly, as an unexpected feature, although a maximum of 2000 reagents were allowed by FINDERS for reaction with REACT2D, more than 2000 were sometimes found and reacted. For instance, in the FINDERS step, Boc-protected amine was kept for reductive amination as FINDERS was instructed to discard amines in the R chains but not carbamates. In the following REACT2D step, *tert*-butyl (4-aminobutyl)carbamate which is a Boc-monoprotected diamine, led to two series of products (Figure 24). On one hand, the free amine (labeled as an amine) was found to be a potential reactive site (X=H) and the corresponding amine was produced with the other amine protected as a Boc (labeled as an amide but not part of the reaction). On the other hand, the Boc-protected amine can be de-protected (X=Boc) and reacted on this site. While after deprotection, both amines can react, one can envision to use orthogonal protection on the free amine. As a result, the two generated products are indeed synthetically accessible.

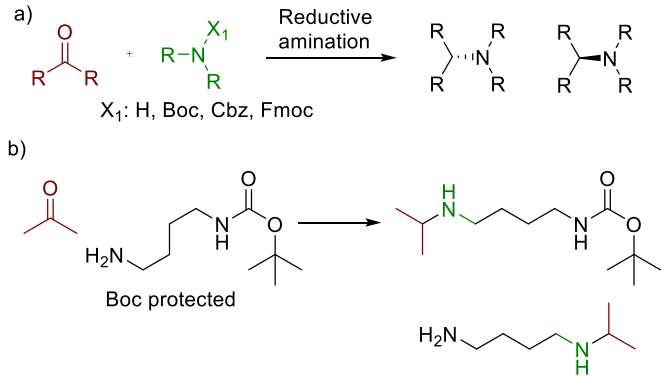


Figure 24. Multiple reaction centers. a) Reductive amination scheme. b) Multiple synthetic outcomes that match the reaction template and are chemically feasible.

Conclusions

Experimentally exploring the synthetically accessible chemical space would require significant resources. As an alternative, computational approaches could be exploited to accurately generate extremely large databases of synthetically feasible compounds in just a few hours. In this context, we have first developed FINDERS, a program which identifies chemicals compatible with a chemical transformation in catalogs and REACT2D, which carries out the combinatorial chemistry necessary to produce the libraries. These two programs are built around substructure search algorithms. Substructure search is important to any laboratory with a digital database. In this regard, the search could be either exact match, to locate information on this molecule for example, or a substructure match, for further work to be done on this reduced library. Through a java interface, which includes a chemical sketcher, chemists can define compatibility rules, draw a chemical transformation, and run FINDERS and REACT2D in just a few clicks. These regulations reduce the returned library and allow the user to define how vast they would like their search; placing R-groups at every possible position will return many structures while, in contrast, one R-group will have fewer hits. An entire protocol of this nature, which puts the control in the user's hand, allows chemists to find synthetically accessible compounds for their desired purpose—chemistry or further computational studies.

ASSOCIATED CONTENT

Supporting Information. The program is available on request from the authors (www.fitted.ca). The sets of reactions (pdf document) and small libraries used for validation (sdf and rxn files) are provided. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
* nicolas.moitessier@mcgill.ca.
‡ current address: Department of Pharmaceutical Chemistry, University of California, San Francisco, 1700 4th St., San Francisco, CA, USA, 94131

ACKNOWLEDGMENT

We thank NSERC (Discovery program) and FRQ-NT (scholarship to JP) for financial support. Calcul Québec and Compute Canada are acknowledged for generous CPU allocations.

EXPERIMENTAL SECTION

The subversion 4832 of the VIRTUAL CHEMIST platform including all the necessary programs used in this work has been used for this study.

References

1. Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H., Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *AAPS J.* **2012**, 14, 133-141.
2. Villoutreix, B. O.; Eudes, R.; Miteva, M. A., Structure-Based Virtual Ligand Screening: Recent Success Stories. *Comb. Chem. High Throughput Screen.* 12, 1000-1016.
3. Matter, H.; Sotriffer, C. Applications and Success Stories in Virtual Screening. In *Virtual Screening*; Wiley-VCH Verlag GmbH & Co. KGaA: 2011, pp 319-358.
4. Irwin, J. J.; Shoichet, B. K., Docking Screens for Novel Ligands Conferring New Biology. *J. Med. Chem.* **2016**, 59, 4103-4120.
5. Sterling, T.; Irwin, J. J., ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, 55, 2324-2337.
6. 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.
7. Xue, L.; Bajorath, J., Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. HTS* **2000** 363-372.
8. Hartenfeller, M.; Schneider, G., De novo drug design. *Methods in molecular biology (Clifton, N.J.)* **2011**, 672, 299-323.
9. Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A., Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Comb. Sci.* **2012**, 14, 579-589.
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. Masek, B. B.; Baker, D. S.; Dorfman, R. J.; DuBrucq, K.; Francis, V. C.; Nagy, S.; Richey, B. L.; Soltanshahi, F., Multistep Reaction Based De Novo Drug Design: Generating Synthetically Feasible Design Ideas. *J. Chem. Inf. Model.* **2016**, 56, 605−620.
12. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating medicinal chemistry, organic/combinatorial chemistry, and computational chemistry for the discovery of selective estrogen receptor

modulatorswith FORECASTER, a novel platform for drug discovery. *J. Chem. Inf. Model.* **2012**, 52, 210-224.
13. Bezanson, M.; Pottel, J.; Bilbeisi, R.; Toumieux, S.; Cueto, M.; Moitessier, N., Stereo- and Regioselective Synthesis of Polysubstituted Chiral 1,4-Oxazepanes. *J. Org. Chem.* **2013**, 78, 872-885.
14. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; 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.
15. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31-36.
16. IUPAC https://iupac.org/who-we-are/divisions/division-details/inchi/.
17. Barnard, J. M., Substructure searching methods: Old and new. *Journal of Chemical Information and Computer Sciences* **1993**, 33, 532-538.
18. Golovin, A.; Henrick, K., Chemical substructure search in SQL. *J. Chem. Inf. Model.* **2009**, 49, 22-27.
19. MDL, In; 2007.
20. Tripos, In; 2009.
21. Crowe, J. E.; Lynch, M. F.; Town, W. G., Analysis of structural characteristics of chemical compounds in a large computer-based file. Part I. Non-cyclic fragments. *J. Chem. Soc. C Org.* **1970**, 990-996.
22. Agrafiotis, D. K.; Lobanov, V. S.; Shemanarev, M.; Rassokhin, D. N.; Izrailev, S.; Jaeger, E. P.; Alex, S.; Farnum, M., Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge. *J. Chem. Inf. Model.* **2011**, 51, 3113-3130.
23. eMolecule, www.emolecules.com. **accessed Oct. 24, 2016**.
24. Kuramochi, M.; Karypis, G. Frequent subgraph discovery. In IEEE, 2001; 2001; pp 313-320.
25. Ehrlich, H. C.; Rarey, M., Searching substructures in fragment spaces. *J. Cheminf.* **2011**, 3.
26. Yu, N.; Bakken, G. A., Efficient exploration of large combinatorial chemistry spaces by monomer-based similarity searching. *J. Chem. Inf. Model.* **2009**, 49, 745-755.
27. Nosrati, G. R.; Houk, K. N., SABER: A computational method for identifying active sites for new reactions. *Prot. Sci.* **2012**, 21, 697-706.
28. Diederich, F.; Kivala, M., All-carbon scaffolds by rational design. *Advanced Materials* **2010**, 22, 803-812.
29. Wang, H., Design of a Structure Search Engine for Chemical Compound Database. *Dissertation, Georgia State University* **2008**, http://scholarworks.gsu.edu/cs_diss/33.
30. *ChemSpider*, Royal Society of Chemistry: 2015.