

# Synthl: a new open-source tool for synthon-based library design

Yuliana Zabolotna<sup>1</sup>, Dmitriy M. Volochnyuk<sup>3,6</sup>, Sergey V. Ryabukhin<sup>4,6</sup>, Kostiantyn Gavrylenko<sup>5,6</sup>, Dragos Horvath<sup>1</sup>, Olga Klimchuk<sup>1</sup>, Olexandr Oksiuta<sup>3,7</sup>, Gilles Marcou<sup>1</sup>, Alexandre Varnek<sup>1,2</sup> \*

**Abstract:** Most of the existing computational tools for library design are focused on the generation, rational selection, and combination of promising structural motifs to form members of the new library. However, the absence of a direct link between the chemical space of the retrosynthetically generated fragments and the pool of available reagents makes such approaches appear as rather theoretical and reality-disconnected. In this context, here we present Synthons Interpreter (*Synthl*), a new open-source toolkit for library design that allows merging those two chemical spaces into a single synthons space. Here synthons are defined as actual fragments with valid valences and special labels, specifying the position and the nature of reactive centers. They can be issued from either the “break-up” of reference compounds according to 38 retrosynthetic rules or real reagents, after leaving groups withdrawal or transformation. Such an approach not only enables the design of synthetically accessible libraries and analogs generation but also facilitates reagents (building blocks) analysis in the medicinal chemistry context. *Synthl* code is publicly available at <https://github.com/Laboratoire-de-Chemoinformatique/Synthl>.

**Keywords:** library design, synthons, fragmentation, enumeration, building blocks, retrosynthesis

## INTRODUCTION

The rational design of chemical libraries for activity screening is crucial for successful drug discovery and

chemoinformaticians have played a highly important role in its rapid development<sup>1</sup> Various computational methods evolved over time to allow chemical data manipulations, structure transformations, de novo generation etc.<sup>2</sup> With such a diversity of existing approaches, the main challenge in modern library design is a trade-off between the theory-inspired novelty introduced by chemoinformaticians and practical considerations of experimentalists.<sup>3</sup> The ability of medicinal chemists to consider both factors is influenced by the availability of the easy-to-use computational tools that provide solutions to the most frequent library design problems while still retaining some level of flexibility embodied in the variety of user-tunable parameters.

Most of the existing technics of de novo library design are based on the generation, rational selection, and combination of promising structural motifs to generate members of the new library<sup>4</sup>. The first task is

1. University of Strasbourg, Laboratoire de Chemoinformatique, 4, rue B. Pascal, Strasbourg 67081 (France) \*e-mail: [varnek@unistra.fr](mailto:varnek@unistra.fr)
2. Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan
3. Institute of Organic Chemistry, National Academy of Sciences of Ukraine, Murmanska Street 5, Kyiv 02660, Ukraine
4. The Institute of High Technologies, Kyiv National Taras Shevchenko University, 64 Volodymyrska Street, Kyiv 01601, Ukraine
5. Research-And-Education ChemBioCenter, National Taras Shevchenko University of Kyiv, Chervonotkatska str., 61, 03022 Kiev, Ukraine
6. Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine
7. Chemspace, Kyiv, Ukraine.

usually achieved by the fragmentation of relevant compounds (for example known ligands of a particular biological target).<sup>5</sup> The resulting fragments or their subset can then be reassembled forming a new library with desired properties. Over the last decades, various methodologies that differ mostly in a set of rules applied for fragment generation and recombination were reported. The most prominent openly available fragmentation method is the retrosynthetic combinatorial analysis procedure (RECAP)<sup>6</sup>. Proposed twenty years ago, it was the first of its kind pseudo-retrosynthetic tool, that applied 11 reaction rules in order to break chemical bonds that can be easily formed via combinatorial chemistry. This methodology together with its latter extension called BRICS<sup>4</sup> has gained extreme popularity and has been used successfully in different drug discovery projects and implemented in several chemoinformatics toolkits, like ChemAxon<sup>7</sup>, OpenEye<sup>8</sup>, and RDKit<sup>9</sup>.

The limitations inherent to the rather small set of reaction rules behind RECAP have been discussed previously, as opposed to the hundreds of automatically extracted reaction schemes introduced in more complex tools for library design and retrosynthetic analysis, like AiZynthFinder<sup>10</sup>, Chematica<sup>11</sup>, ICSYNTH<sup>12</sup> etc. It is usually claimed that such tools are covering the scope of known chemical reactions more comprehensively. On the one hand, they indeed reflect up-to-date synthesis expertise, but at the same time, they include some sophisticated protocols pertaining to synthetic creativity, rather than an optimal solution for everyday routine problems. Considering how uncertain is the success of the drug design campaign at its early stages, investing more time and resources in the synthesis of the initial screening libraries does not seem very efficient. Therefore, medicinal chemists traditionally use only a tiny fraction of the reactions that allow faster advancement in drug discovery projects, saving complex elaborated procedures for optimization of confirmed leads<sup>13-16</sup>.

This tendency is advocated in a recent study, showing that molecular quality, comprising molecular complexity, diversity, and novelty, is typically not related to the type of chemical reactions used to produce screening compounds (excepting targets for which only natural product-like ligands are known).<sup>17</sup> Their diversity, complexity, and novelty are more influenced by the quality of the selected building blocks (BBs). In this context, the absence of the direct link between the chemical space of the generated fragments and the pool of available BBs makes tools

like RECAP and BRICS appear as rather theoretical, reality-disconnected approaches, distant from down-to-earth practical library design based on the reagents present in the laboratory drawers.<sup>18</sup> Some methodologies of library de-novo designs considering BBs availability have been previously reported, including both commercial/proprietary software<sup>19-21</sup> and methodologies used mostly by the authoring academic group<sup>22</sup>.

Here we describe a new open-source toolkit for synthons-based library design, called Synthons Interpreter (SynthI). In chemoinformatics synthons were first introduced by R.D.Cramer et al.<sup>23</sup> in 2007 as structures with one or more open valences each having a defined reactivity. In this work, synthons are defined differently: the open valence at the connection/disconnection point is complemented by hydrogen atom(s) and a special label determining its reactivity is assigned. The label is associated with those reagent classes (in total there are almost 150 mono- bi- and trifunctional subclasses) that can produce a given synthon (see Table 1). Their chemical validity allows to treat synthons as any other chemical structures: to assess different properties using machine-learning models, to evaluate similarity, and to visualize their chemical space. In the unified scheme presented here synthons can be transparently issued from either the “break-up” of reference compounds according to 38 pseudo-retrosynthetic rules, or from real reagents, after leaving/protective groups removal or any other transformations required to generate the moiety inherited by the reaction product. As a result, SynthI can be used for several tasks: i) analysis of the available BBs collections; ii) global enumeration of all compatible synthons combinations based on the selected reactions and available BBs; iii) detection of BBs producing synthons that are needed to synthesize desired compounds; iv) synthons-based focused library design – a combination of synthons identical or analogous to those obtained via pseudo-retrosynthetic fragmentation of active compounds.

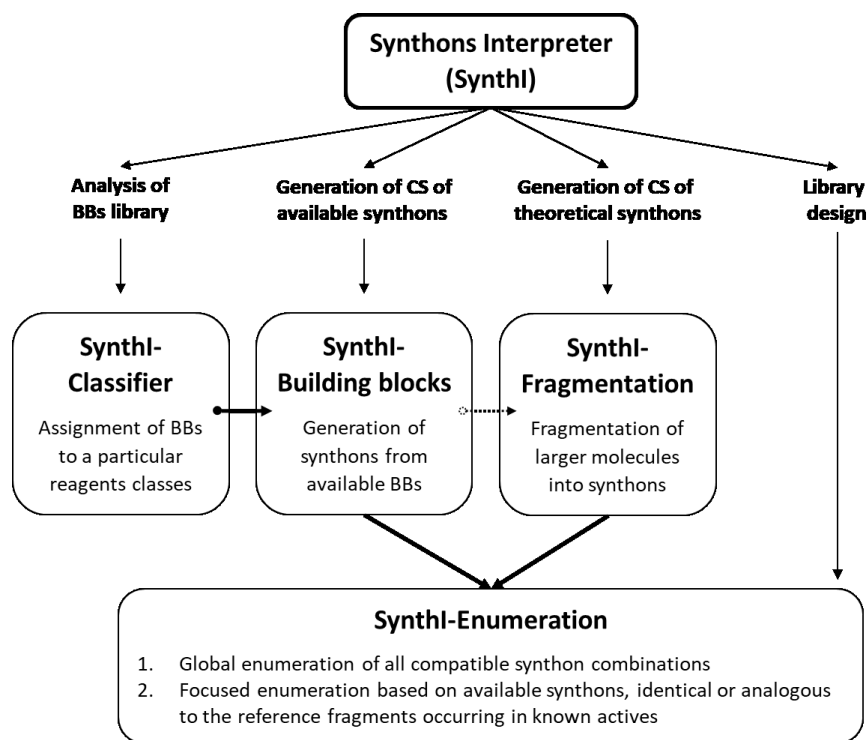
## IMPLEMENTATION

### General description of SynthI

SynthI is a python3 RDkit-based<sup>9</sup> (2021) library that generates synthons from larger molecules via fragmentation or from small reagents via functional group transformations. Being a knowledge-based tool, SynthI is based on the extensive library of SMARTS, defining each reagent class and SMIRKS that specify the reaction rules for synthon generation from BBs,

pseudo-retrosynthetic bond disconnections, or synthon recombination. SynthI consists of four modules (Figure 1), each being responsible for a

particular task. In the following chapters, you can find a detailed description of each of them.



**Figure 1.** SynthI functionality: analysis of BBs libraries, achieved with SynthI-Classifier; generation of chemical space (CS) of available synthons from the BBs after their classification - SynthI-BBs; generation theoretical synthons CS via fragmentation of larger compounds (with or without the use of available synthons library for prioritizing the fragmentation schemes resulting in a higher portion of available synthons) – SynthI-Fragmentation; library design via global or focused enumeration – SynthI-Enumeration.

### SynthI-BBClassifier

The first step in BB processing is a selection – a binary decision-making algorithm returning whether a given molecule may or may not qualify as a reagent of a specified class in a specified reaction. This involves three key aspects:

- Detection of the required characteristic functional group[s] characterizing the envisaged reagent class, which can straightforwardly be achieved by SMARTS pattern matching.
- Analysis of the chemical context in which the characteristic functional group is placed, and which modulates its reactivity. This is a weak point of the procedure because these effects are often long-range (conjugation, inductive effects), geometry-dependent (steric effects, intramolecular hydrogen bonds) and, of course, overlapping (several substituents inducing conflicting and not always additive effects). In absence of a robust global model of chemical

reactivity, SMARTS encoding of the most often seen and impactful structural patterns associated to a loss of functional group reactivity is the only practical solution so far.

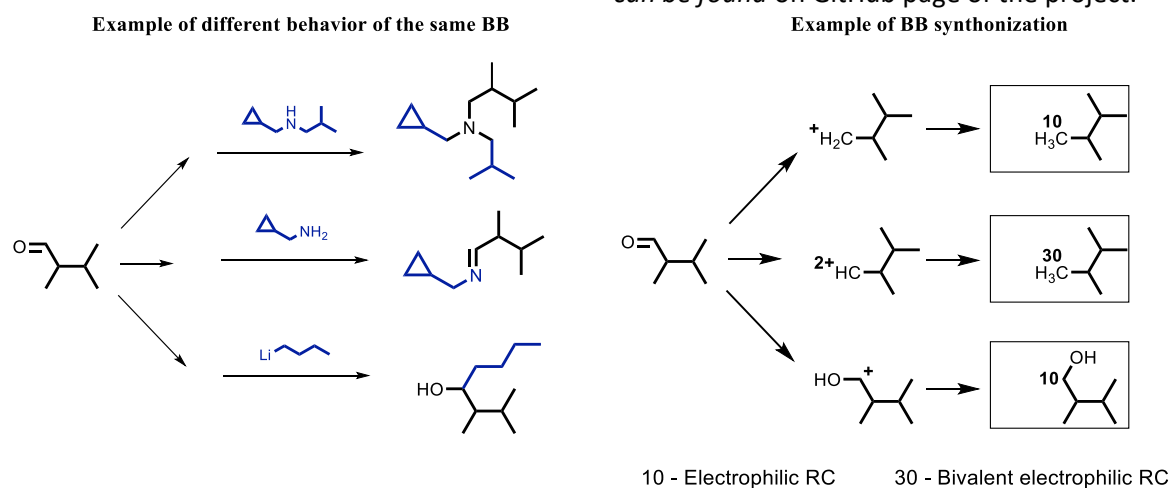
- Detection of unprotected competing or cross-reacting functional groups, likely to trigger secondary reactions leading to a mixture of products. For example, in order to be effectively used as an aldehyde reagent, BB should not contain structural moieties of acylators, alkylators, unprotected amino groups, thiols, isocyanates, metalorganics, etc. These may also be provided as a list of SMARTS patterns.

The full list of SMARTS for the BBs classification is provided in *SMARTSlib.json* and *SynthI\_AllSmartsFromClassifier.xlsx* files on GitHub page. In total, 22 monofunctional BB classes were considered, like acyl halides, boronics, ketones, primary amines etc. Almost each of them incorporates subclasses, totaling up to 100. For example, class “Alcohols” includes three subclasses that would have different reactivity – “Heterols”, “Aliphatic alcohols”

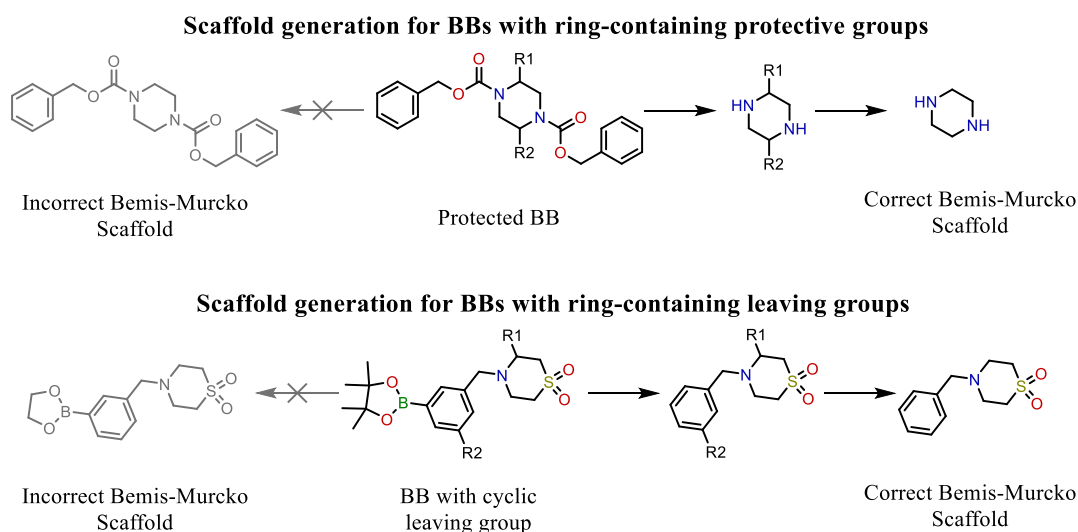
and “Phenols”. In addition, there are 28 bifunctional and 19 trifunctional classes. All of them concern only reagents for coupling reactions as soon as the given version of Synthl does not include heterocyclization reactions. From the library design point of view, their usage would lead to the destruction of the privileged scaffolds that contribute significantly to the exhibited activity. Therefore, in the first implementation of Synthl heterocyclization reactions were not taken into consideration. For more detailed retrosynthesis, however, those reactions are highly important, therefore we are currently working on the implementation of the Synthl-Heterocyclization module, that would allow the user to select whether they want to include cycle bonds disconnection.

### Synthl-BBs

The same BB can be assigned to several classes followed by the generation of synthons, corresponding to each class using Synthl-BBs module. In each synthon, the special labels are placed at the former position of the leaving groups (**Figure 2** and **Table 1**). They define the type of the bond disconnection and reaction center (RC) – electrophile, nucleophile, radical, etc. The full list of unique synthons generated from the user-provided BBs library produces a chemical space of available synthons. In the case of a compound, containing protective groups it is up to the user to decide whether to keep protected synthons or not (*keepPG* option). The list of all synthons generated from each BB class is provided in *Synthl\_BB\_classes\_and\_respectiveSynthons.xls*, which can be found on GitHub page of the project.

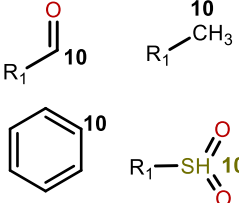
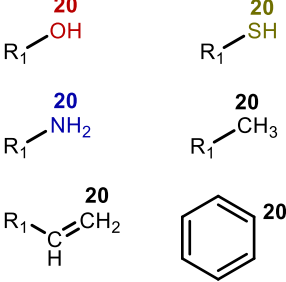

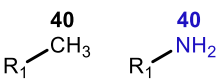

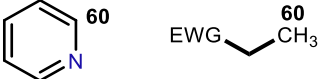
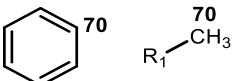
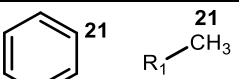
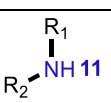


**Figure 2.** Example of different behavior of the same BB (here - aldehyde) and generation of corresponding synthons. Labels on the synthons define the nature of the reaction center (RC).



**Scheme 1.** Scaffold generation in BBs analysis. Ring-containing protective and leaving groups should be removed before generating a scaffold.

**Table 1.** Synthons labels and examples of corresponding reagents.

Synthon Label	Examples of Synthon	Nature of the reaction center (RC)	Example of corresponding reagent classes
AH <sub>n</sub> :10		Electrophilic	Acyl, aryl and alkyl halides, sulfonylhalides, anhydrides, acids, aminoacids, esters, alcohols, aldehydes, ketones, Weinreb amides, acylated azides, iso(thio)cyanates, oxiranes
AH <sub>n</sub> :20		Nucleophilic	Alcohols, thiols, amines, amides, NH-azoles, hydrazines, hydrazides, hydroxylamines, oximes, esters, element organics, metal organics, ketones, aryl and allyl sulphones, alkenes for Heck couplings
CH <sub>n</sub> :30		Bivalent electrophilic	Aldehydes, ketones
AH <sub>n</sub> :40		Bivalent nucleophilic	Ketones, primary amines, hydrazines, hydroxylamines, reagents for olefination (Julia-Kocienski, Wittig, Horner-Wadsworth-Emmons)
CH <sub>3</sub> :50		Bivalent neutral	Terminal alkenes (for metathesis)
CH <sub>n</sub> :60		Electrophilic radical	Minisci CH-partners, Michael acceptors
CH <sub>n</sub> :70		Nucleophilic radical	BF <sub>3</sub> and MIDA boronates, oxalate alkyl esters, NOPhtal alkyl esters, sulphinates
CH <sub>n</sub> :21		Boronics-derived nucleophilic	Boronic reagents
NH:11		Electrophilic nitrogen	Benzoyl O-acylated hydroxylamines

### Scaffold generation for BBs

The most common approach for the structural analysis of any compound library is to generate scaffolds<sup>24</sup> - cyclic molecular cores without side chains - and count the frequency of their occurrence in the compound collection<sup>25</sup>. For the analysis of reagent libraries, BB structures need to be preprocessed prior to the scaffolds generation by removing any ring-containing moieties that are not parts that will not be kept in the reaction product and thus are irrelevant in BB analysis (**Scheme 1**). It includes some protective (benzyl (Bnz),

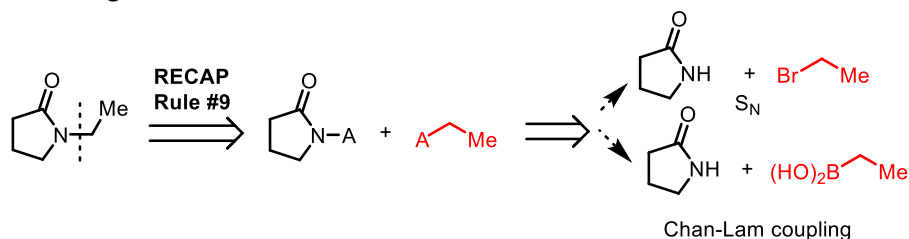
benzyl carbamate (Cbz) and fluorenylmethyloxycarbonyl (Fmoc)) and leaving groups (boronics, oxiranes). Based on such preprocessing, SynthI allows to generate relevant BBs scaffolds, count their occurrence in the provided collection of BBs, and even construct cumulative scaffold frequency plot.

### SynthI-Fragmentation

The chemical space of theoretically relevant synthons can be generated via pseudo-retrosynthetic bond disconnection of the relevant compounds (e.g. ligands

of a particular target) implemented in SynthI-Fragmentation. It is based on the most common combinatorial reactions, expressed via SMIRKS. Previously, 11 RECAP bond cleavage rules were proposed based on the “commonly used” combinatorial chemistry. However, after more than 20 years these rules needed to be revised in accordance with modern synthetic techniques. In addition, in RECAP and BRICS for each type of bonds there was only one disconnection rule. However, the same bond can be formed by different reagents via reactions that can

have completely different mechanisms. For example, N-alkylation of lactams can be performed via nucleophilic substitution of alkyl halides or via Chan-Lam coupling with boronic acids (Scheme 2). In this context, in order to be able to link the chemical space of available synthons, generated from provided BBs library, to the synthons resulted from fragmentation, several rules of disconnection are needed for the same bond type.



**Scheme 2.** Example of RECAP disconnection of the bond that can be formed via different reactions.

The reaction rules behind SynthI were collected based on the analysis of current literature and our experience in medicinal chemistry synthesis. It included various reactions, leading to:

- several ways of disconnection of the same strategic bonds that were already considered in RECAP and/or BRICS (Buchwald-Hartwig amination<sup>26</sup>, Cu-mediated C-N/O coupling<sup>27</sup>, umpolung cross-coupling<sup>28</sup>, Chan-Evans-Lam coupling<sup>29</sup>, olefin metathesis<sup>30</sup>, non-classical carbonyl olefination (like Julia-Kocienski)<sup>31, 32</sup>, C-H activation<sup>33</sup>, sulfonyl fluorides chemistry<sup>34</sup>, Suzuki  $C_{Ar}-C_{sp3}$  cross-coupling, novel methods for  $C_{Ar}-C_{sp3}$  couplings).
- disconnection of the new strategic bonds absent in the previous implementation (Heck  $C_{Ar}-C_{sp2}$ , Sonogashira  $C_{Ar}-C_{sp}$  and Suzuki  $C_{sp2}-C_{sp2}$  couplings, imines, oximes, hydrazones and semicarbazones synthesis, sulphinic acid salts alkylation and their Cu-catalyzed arylation)

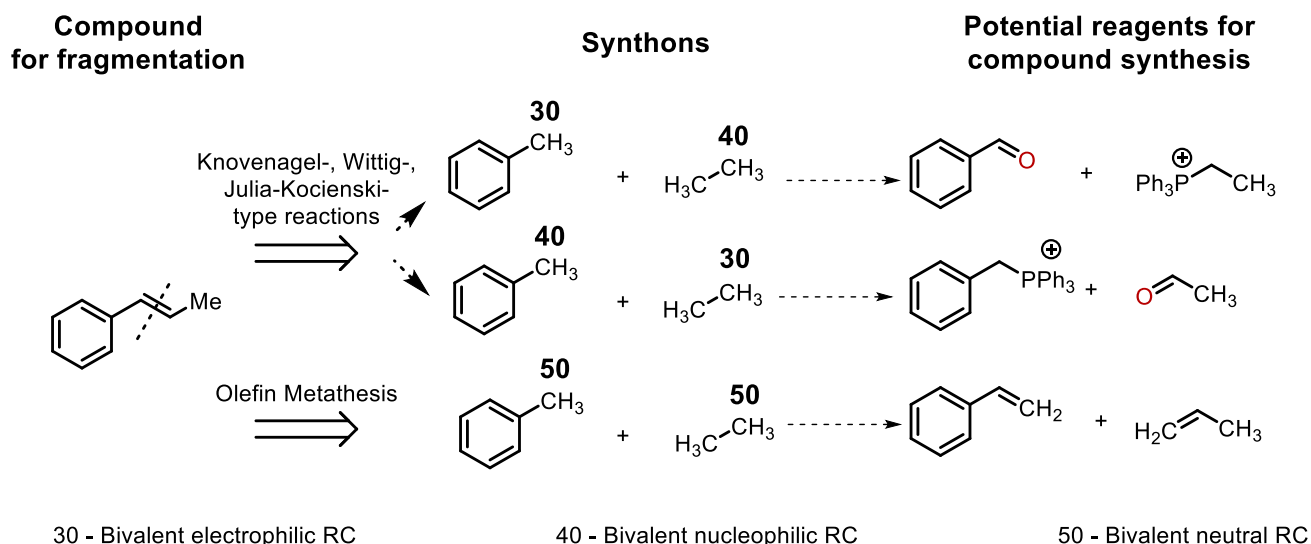
Also, the set of new radical chemistry, as well as new methods of late-stage functionalization (Baran diversinates<sup>35</sup>, Minisci-type reaction<sup>36</sup>), were included in SynthI. These new reactions dramatically changed modern retrosynthetic thinking of the medicinal chemist<sup>14, 37</sup>, and the new more effective conditions for such reactions still actively investigating<sup>38</sup>.

In total SynthI contains 13 broad reaction types for the bond disconnections and 37 subtypes, that may lead to different synthons. For example, for the “Olefination” type, there are two subtypes – “Knoevenagel-, Wittig-, Julia-Kocienski- type reactions” and “Olefin Metathesis”. The first one is the example of polar bond disconnection resulting in bivalent electrophilic and nucleophilic synthons, while the second one produces neutral biradicals (**Scheme 3**). Obtained synthons can be traced back to the potential BBs for compound synthesis. The full list of reaction rules with some examples is available in the Supporting Information.

SynthI-Fragmentation allows one to select a subset of reactions, but in this study, all of them are used. After each cut, the combination of synthons from which molecule can be synthesized is stored. If more than one bond in a molecule can be disconnected, then the hierarchy of all possible disconnections and resulting synthons combinations are stored. Given the list of “available” synthons provided by the available BBs, fragmentation schemes predominantly returning fragments listed amongst these available synthons are obviously preferable. The availability rate is herein defined as the percentage of heavy atoms of the fragmented compound that can be provided by available synthons:

$$\text{Availability rate} = \frac{\sum \text{heavy atoms coming from available synthons}}{\text{Total number of heavy atoms in a molecule}} * 100\%$$





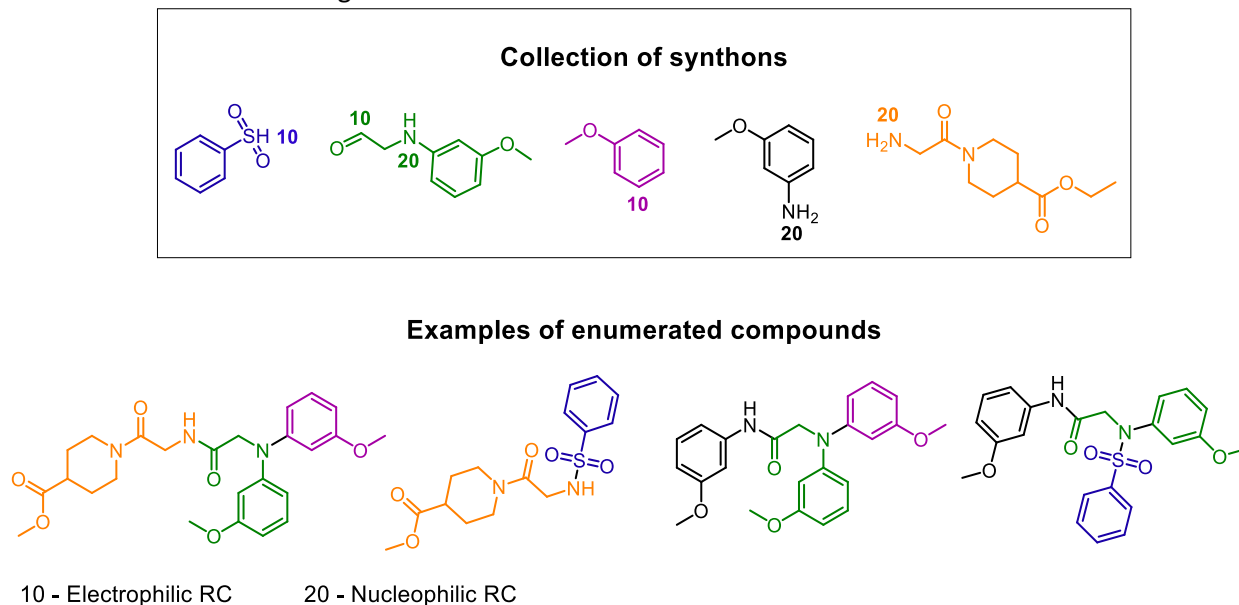
**Scheme 3.** Example of Synthl reaction type with two subtypes representing different mechanisms of the same bond formation/disconnection. Labels on the synthons define the nature of the reaction center (RC).

Based on this value, the optimal pathway can be selected to be written into the summary output file (see SI). One may also navigate the disconnection hierarchy using several built-in functions. More details on the usage of Synthl and tutorial can be found on the GitHub page (<https://github.com/Laboratoire-de-Chemoinformatique/Synthl>).

combinatorial library of all compounds that can be synthesized using a given set of synthons (**Figure 3**). Users can control the maximum number of synthons that can be combined together. As well as the list of reactions for enumeration. If the maximal number of synthons has been reached but some open RCs were left this product will be discarded.

### Synthl-Enumeration

This last module applies the list of the abovementioned reaction rules in order to generate the full



**Figure 3.** Example of library enumeration using a user-provided collection of synthons.

Synthl-Enumeration also allows to generate a focused library of the synthesizable analogs of the provided compound. The input molecule is first

fragmented up to the smallest synthons. Their availability is checked using the BBs synthons library. The same library is used for the search of the analogs

of generated synthons - synthons containing the same types of RCs (but not necessarily in the same positions), the same number of rings, and matching the constraints, adopted from the positional analogs scanning (PAS) strategy for lead optimization<sup>39</sup>. According to the latter, analogs should be a substructure/superstructure of the original compound (in our case synthon) and differ from it only in the absence/presence of one functional group: CH<sub>3</sub>, F, NH<sub>2</sub>, OH or be a result of C<sub>Ar</sub>->N<sub>Ar</sub> or N<sub>Ar</sub>->C<sub>Ar</sub> replacements. These rules have been changed slightly to soften the criteria for synthons selection in order to enable producing more comprehensive focused libraries. Thus, the structural isomers were also considered analogs. In addition, there is a possibility for the user to specify the synthons similarity threshold that will be applied independently of the previous filters for the search of additional analogs of the original synthon via similarity approach. The rules concerning RC types and number of rings are used for all analogs selection including similarity. The Tanimoto coefficient is calculated with RDKit using Morgan fingerprints (radius=2, nBits=2048) as descriptors.

With *strictAvailabilityMode* only synthons that were found in the available BBs or have available analogs are selected for library generation. If one of the required synthons does not have any direct or analogous correspondence in the provided BB library, easily synthesizable analogs for the input molecule can not be generated. Otherwise, unavailable synthons will be also used for focused library design. The new library generation is based only on the reaction according to which compound was fragmented. The number of combined synthons is fixed to the number of synthons obtained via molecule fragmentation in a selected synthetic path.

## DATA FOR CASE STUDY

As a source of available BBs, the library of 201 675 in-stock reagents provided by Enamine was used. 79 drugs, recently approved by FDA have been used as a dataset for fragmentation and analogs generation. The full list together with fragmentation results can be found in Supporting Information.

## RESULTS AND DISCUSSION

The weak spot of the RECAP-like tools is their potentially low propensity to propose the exact same fragments that are provided by real-world BBs ready to use in the laboratory. This gap can be bridged by

introducing an unified chemoinformatics formalism to handle the synthon chemical space of both RECAP fragments and BB-provided, "available" synthons. From one point of view, the nature of BB is determined by the protected and unprotected reactive functional groups it contains. They define the list of reactions BB can participate in, and partners it can react with. However, in the medicinal chemistry context, those leaving groups are less interesting than the structural, pharmacophoric or physico-chemical features that will be contributed by the BB to the final molecule. One BB, used under different conditions can contribute differently to the final molecule, while the same structural fragments can be introduced by different BBs (**Figure 2**). Using synthons as a unified representation, SynthI allows merging the chemical space of BBs (or rather structural increments that they bring to the final molecule) with a chemical space of fragments, obtained via pseudo-retrosynthetic bond disconnections. The herein-developed system of labels encodes the position and chemical nature of the reactive centers while preserving structure validity, allowing to treat synthons as actual compounds. This not only enables the design of synthetically accessible libraries but also facilitates BB analysis in the medicinal chemistry context.

### BB classification, synthonization and scaffold analysis

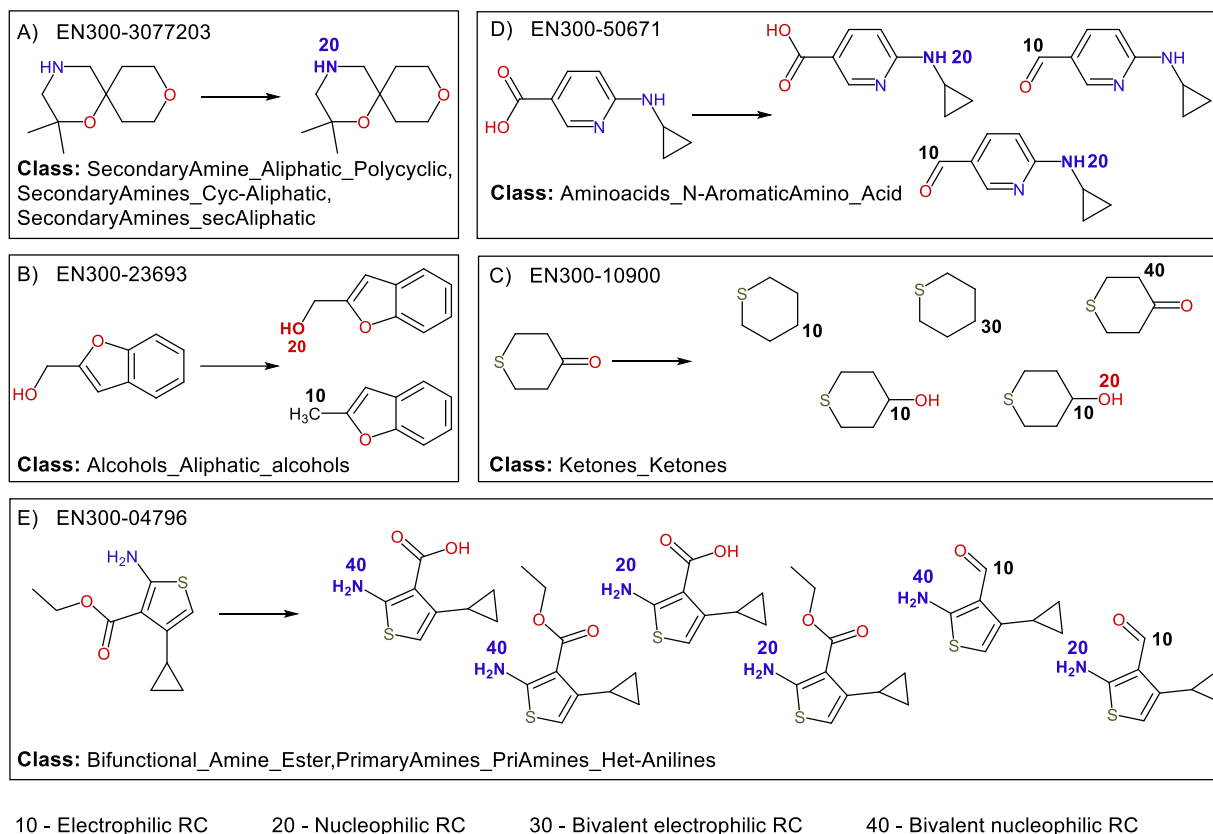
Out of 201 675 BBs used in this work, 18 were not processed by RDKit and 25 414 reagents were not assigned to any classes implemented in the first version of SynthI (mostly reagents for heterocyclization like nitriles, oximes, etc.). For the remaining 176 261 BBs, 388 019 synthons were generated. In **Figure 4** one can see examples of BB classification and synthonization. Some of the BB classes, e.g. secondary amines, produce only one synthon per BB (**Figure 4A**). Others, like ketones, can result in numerous synthons depending on the reaction conditions (**Figure 4C**). An example of aminoesters synthonization with option *keepPG* is shown in **Figure 4E**.

The advantage of adopted synthon representation is that in SynthI synthons are neutral structures with valid valences. The RC position and nature are encoded via atom mapping, which does not change the synthon structure. This allows to analyze them as any other compounds. For example, it is possible to calculate their physicochemical properties and filter them according to the rule of two (Ro2). This rule has been introduced by Goldberg et al.<sup>40</sup> as a simple way of BBs prioritization for designing compounds with physical properties that are suitable for oral administration.

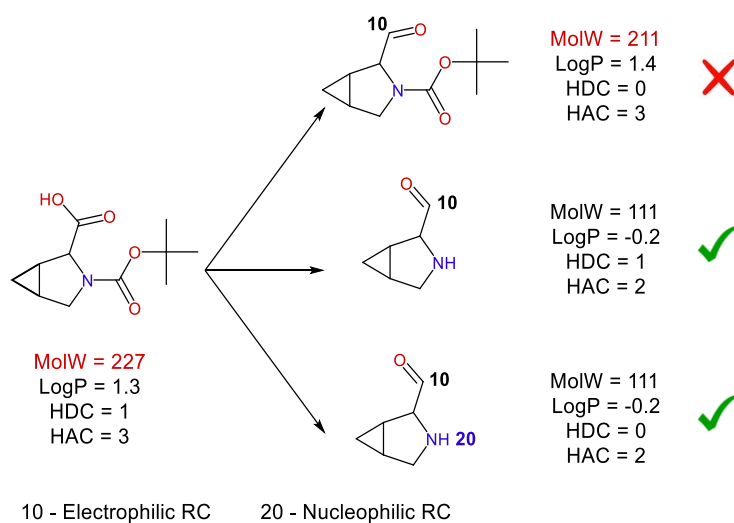


According to Ro2, increment that will be introduced to the molecule by BB should have such properties: MW  $\leq 200$ , logP  $\leq 2$ , H-bond donors  $\leq 2$ , H-bond acceptors  $\leq 4$ . SynthI allows filtration of synthons according to this rule at the stage of synthons library generation

from available BBs, fragmentation (for the synthesability check) or analogs library enumeration (for control of the physical properties of generated compounds) (Figure 5).



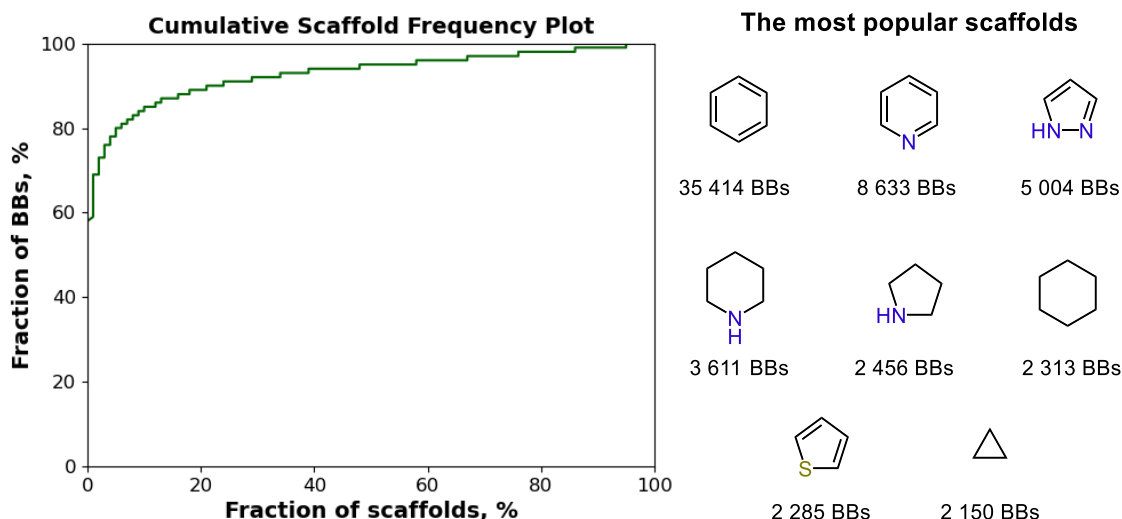
**Figure 4.** Examples of BB classification and synthonization. Labels on the synthons define the nature of the reaction center (RC).



**Figure 5.** Ro2 synthons filtering for BB prioritization (MW  $\leq 200$ , logP  $\leq 2$ , H-bond donors  $\leq 2$ , H-bond acceptors  $\leq 4$ ).

Scaffoldization of 200K Enamine BBs resulted in 19 820 scaffolds with the majority of them (12 272 or 62%) being singletons (occur only in one BB). As one can see in **Figure 6**, a very tiny fraction of scaffolds (<1%) covers almost 60% of BBs from the analyzed collection. The

most frequent scaffolds are simple one-ring structures - benzene, pyridine, pyrazole, piperidine, pyrrolidine, cyclohexane, thiophene, and cyclopropane – and the diversity of BBs libraries is mostly gained via their side chains decorations.



**Figure 6.** Scaffold analysis of the BBs library.

### Fragmentation of FDA approved drugs

As a case study for SynthI-Fragmentation, 79 drugs FDA-approved in 2020 were used examples of compounds to be circumscribed by focused combinatorial libraries of analogues, using the above-processed available BBs. All molecules, except osilodrostat, were fragmented and the optimal set of 2-6 synthons were selected. Out of them, 8 molecules resulted in a set of synthons with a 100% availability rate (all required synthons were incarnated in existing BBs). In order to evaluate the accuracy of the proposed fragmentation schemes from the experimental synthesis perspective, it was compared to the published synthetic pathways (found using Reaxis<sup>®41, 42</sup> and SciFinder) for each of the case study drugs (see Supporting Information). For 24 drugs, SynthI fragmentation fits perfectly to the experimentally validated synthetic procedures. Fragmentation results for the other 18 drugs have minor discrepancies caused by the absence of heterocyclization and reduction/oxidation reactions. Heterocyclization reactions prevail in the synthesis of the remaining compounds and thus corresponding literature data for these compounds cannot be fairly compared to SynthI fragmentation results.

In **Scheme 4** one can see the hierarchy of synthons and reactions, resulted from the fragmentation of cenobamate. SynthI-Fragmentation produced four

synthetic pathways, each including two stages. The optimal pathway consisted of consecutive application of  $S_N$  alkylation and O-acylation disconnection rules. Two out of three resulted synthons were found in the provided synthons library (availability rate = 72%). The synthetic pathway found in literature is highly similar to the one, proposed by SynthI<sup>43</sup>. The difference is in the usage of the 2-bromo-1-(2-chlorophenyl)ethanone as a precursor for 2-bromo-1-(2-chlorophenyl)ethanol and chlorosulfonyl isocyanate instead of trichloroacetyl isocyanate for the introduction of carbamate moiety.

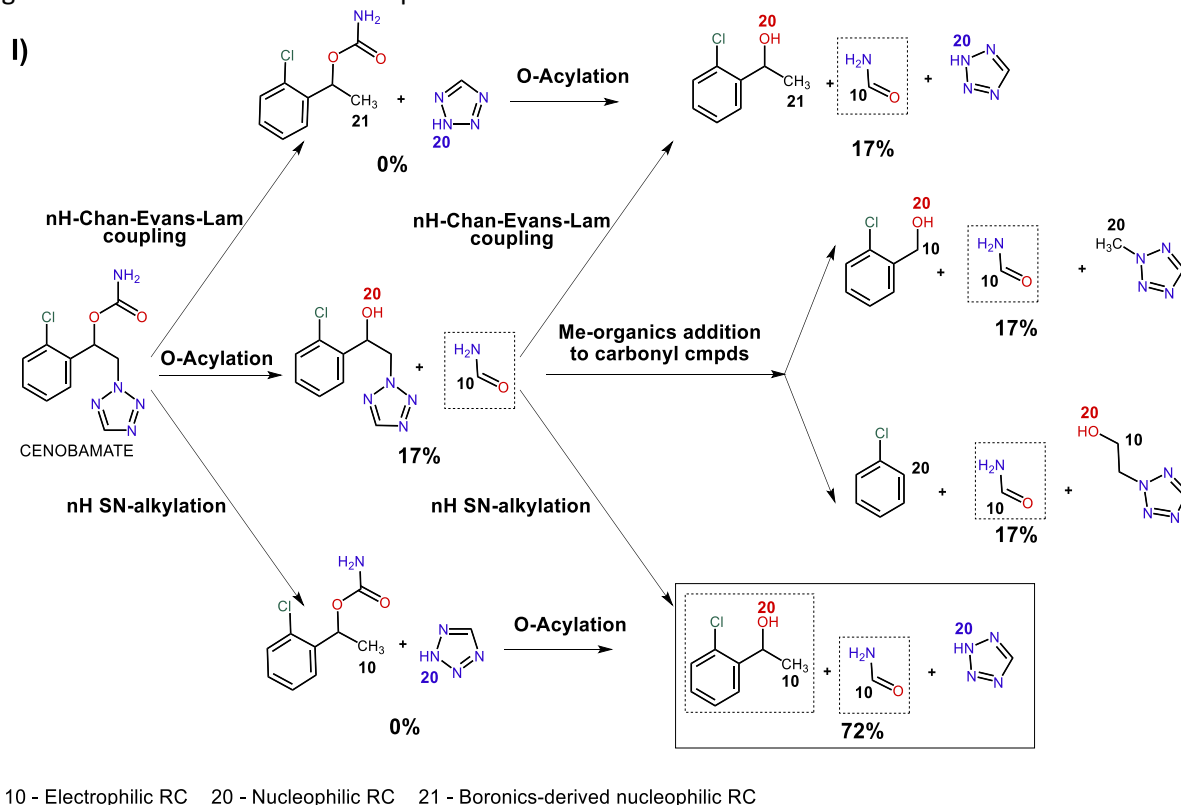
### Analog search case study

Exploring analogs of a reference molecule in terms of combinations of analogues of its constituent BBs is widely used for navigation of very large commercial and proprietary virtual libraries like WuXi Apptec, Enamine REAL (1.3B)<sup>44</sup>, Enamine REAL space (29B)<sup>45</sup>, Eli Lilly PLC (10<sup>10</sup>)<sup>46</sup>, BICLAIM by Boehringer Ingelheim (10<sup>11</sup>)<sup>47</sup>, Pfizer Global Virtual Library (10<sup>14</sup>)<sup>21</sup> etc. All of them are based on the fixed internal collections of reagents and reactions, but with the help of SynthI, it becomes possible to navigate in a similar manner a customized non-combinatorial chemical space, defined by the user-selected reactions and BB collections.

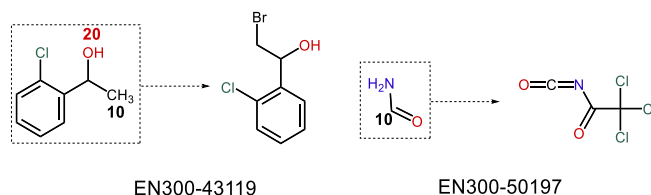
With the help of SynthI, one can perform a retrosynthetic fragmentation of compounds of interest, search for the available BBs producing synthons that

are similar to the resulting fragments, and thereupon enumerate analogs of the initial compound. As a result of Synthl application with activated *strictAvailabilityMode* and additional similarity synthons selection option (with Tanimoto coefficient  $\geq 0.5$ ), analogs for 23 out of total 79 drug compounds were generated. The number of compounds in the

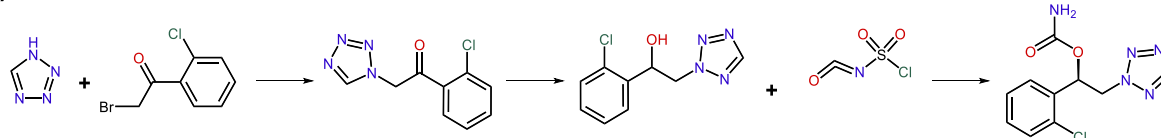
analog libraries varies significantly - from 4 compounds for cenobamate to almost 7M for fedratinib (see Supporting Information). The size of the analog libraries depends on the number of synthons resulted from initial compound fragmentation and the number of analogs synthons found in the Enamine collection.



II)



III)



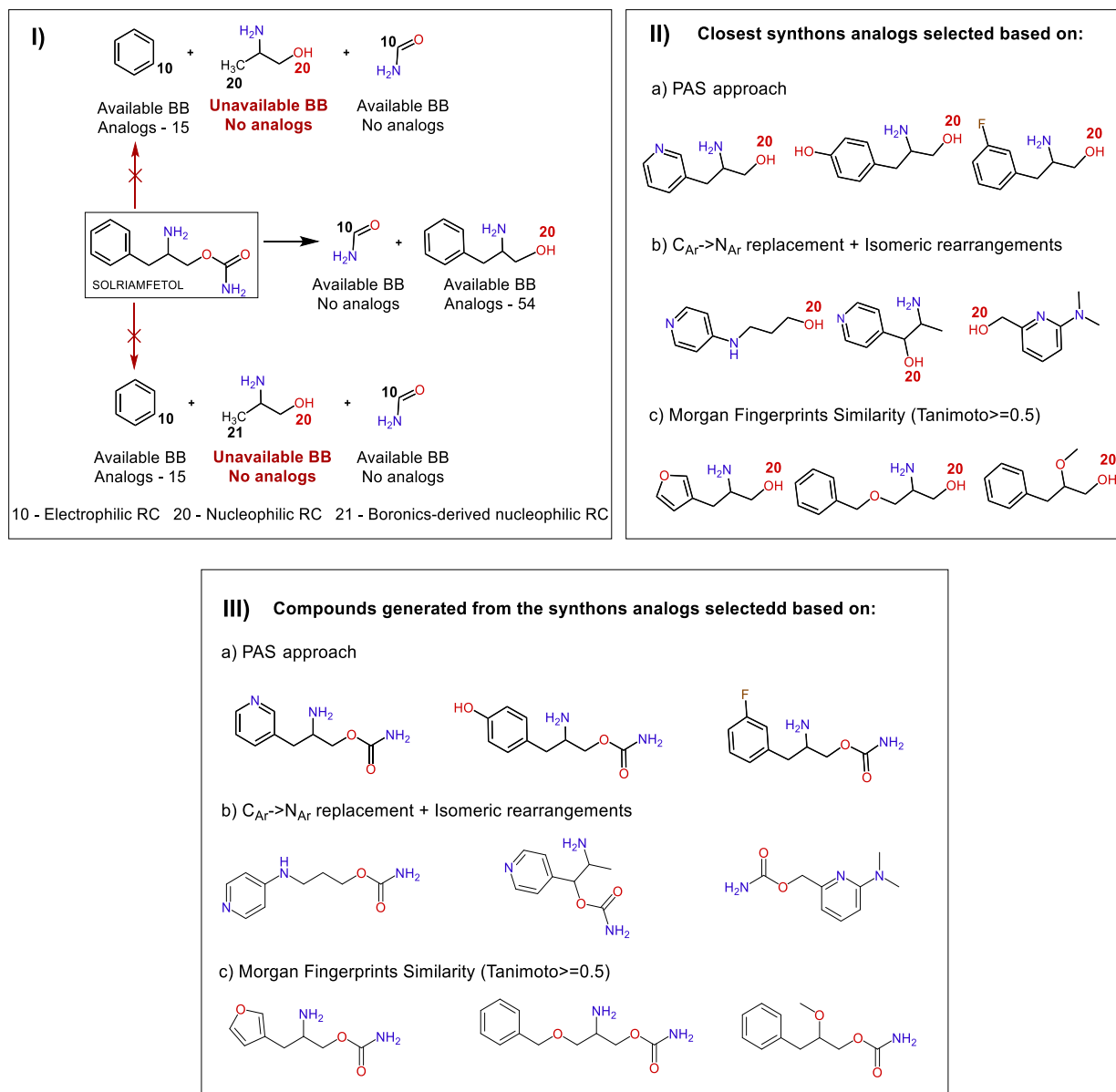
**Scheme 4.** (I) Example of Synthl fragmentation of cenobamate with the full synthetic hierarchy and experimentally validated synthesis of this compound. The number near the selected set of synthons corresponds to its Availability Rate, %. (II) Available BBs, their identifiers in Enamine catalog and related synthones (in dashed frames). (III) Synthesis of cenobamate reported in reference<sup>43</sup>.

In **Figure 7** one can see an example of the analog generation for solriamfetol. For this molecule, there are three possible fragmentation schemes, but only

one of them results in a set of synthons that are present as such or represented by close analogs in the available synthons library. As it was previously

explained in the methods, there are several sets of rules according to which two synthons may be considered analogs: i) they differ by simplest PAS modifications, ii) are isomers of each other or iii) have synthon similarity above a specified threshold (here Tanimoto coefficient  $\geq 0.5$ ). In **Figure 7** the examples of

synthon analogs for each of these categories are given. Solriamfetol analogs generated using them are also provided and as one can see, they are structurally very close to the starting drug, but still providing some level of diversity inside the focused solriamfetol library.



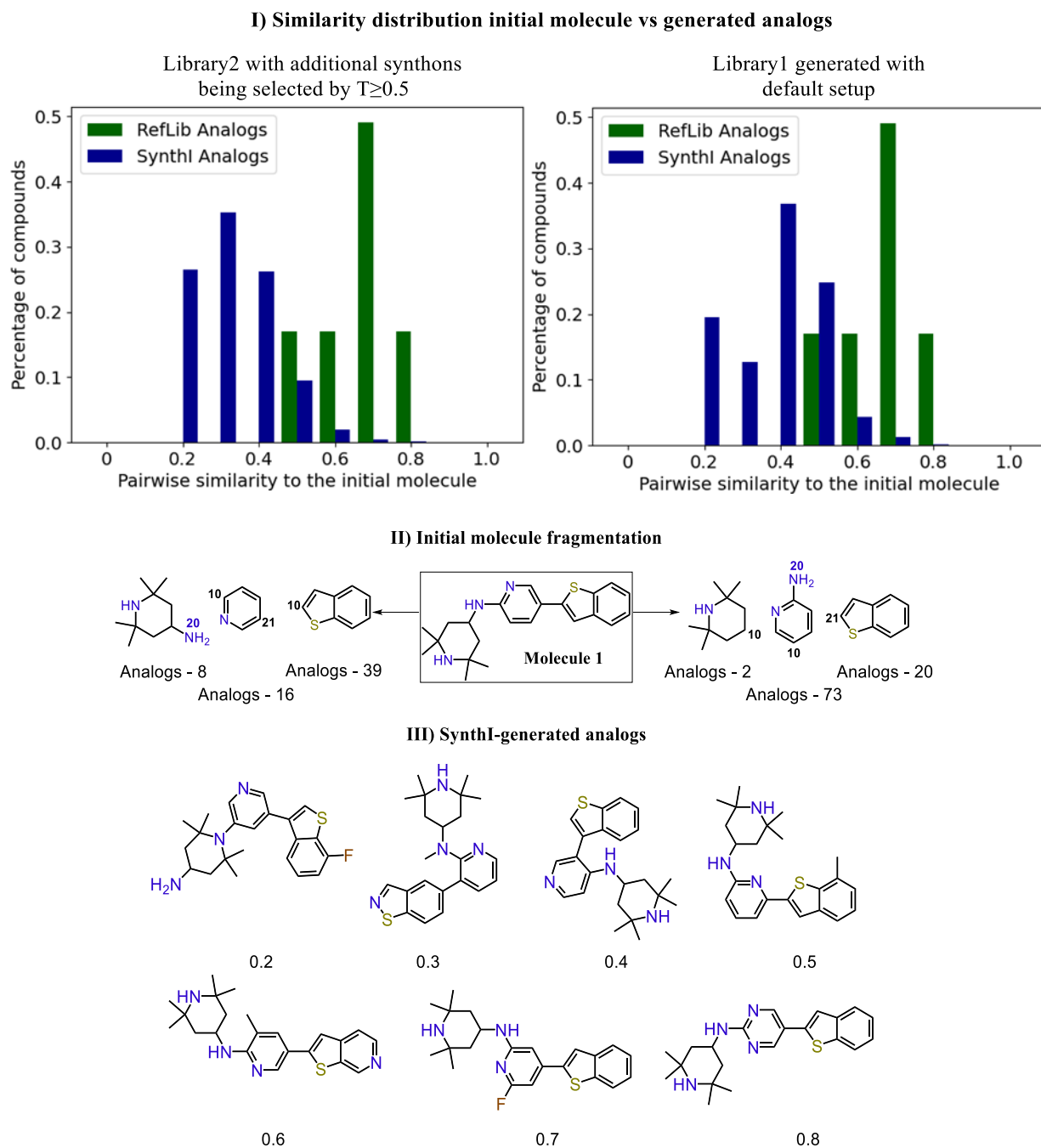
**Figure 7.** Synthons-based generation of solriamfetol analogs. (I) Solriamfetol fragmentation and synthetic pathways selection. (II) Selection of the closest synthon analogues based on (a) PAS approach, (b) C<sub>Ar</sub>->N<sub>Ar</sub> replacement + isomeric rearrangements and (c) Morgan Fingerprints Similarity (Tanimoto  $\geq 0.5$ ). (III) Compounds generated from synthons selected at the step (II).

Considering that the similarity score is always a function of selected descriptors, for the unbiased analysis we need the reference library that would serve as some kind of internal “calibration” scale of the similarity score. In order to create such a library, the simplest PAS modifications (CH<sub>Ar</sub>→F, CH<sub>Ar</sub>→OH, CH<sub>Ar</sub>→CH<sub>3</sub>, CH<sub>Ar</sub>→NH<sub>2</sub> and CH<sub>Ar</sub>→N<sub>Ar</sub>) of the chemical

structure of the reference compound (Molecule 1 **Figure 8**) was performed. Note that modifications were applied manually to the whole structure of the reference compound and not to the underlying fragments like it is done in SynthI. As a result, the reference focused library (RefLib) containing 53 analogs of Molecule 1 was obtained. These compounds

differ only by one atom from the reference molecule, thus their similarity to it can set up a “baseline” of what

to consider as similar compounds in the chosen descriptor space.



**Figure 8.** I) Comparison of the similarity distribution between the initial molecule and three analogs libraries (RefLib, SynthI-generated Library1 (default setup) and Library2 (additional synthons being selected by the  $T \geq 0.5$ ). II) Fragmentation of the initial Molecule1 and number of analogs found for each synthon. III) Examples of generated analogs of Molecule1 with different similarities to the initial compound. The numbers correspond to pairwise Tanimoto similarity with Molecule 1.

From the other side, with the help of SynthI-Enumeration we have generated two libraries of analogs: i) Library1 - 2 593 compounds with a default SynthI setup and ii) Library2 - 8 928 compounds with

activated similarity synthons selection (additional synthons were selected as analogs if their similarity to one of the original synthons was higher than 0.5). Morgan Fingerprint similarity between Molecule 1 and

each member of these two libraries was compared to the same values for the 53 closest analogs from RefLib. As one can see from **Figure 8 (I)**, Synthl-generated compounds, especially from Library2, possess higher diversity with respect to Molecule 1 than analogs from RefLib. This is an expected and desired result, that follows from the adopted approach of the search of synthons analogs rather than direct analogs of the molecule. In the second case, only a single modification is allowed for the whole molecule, while in the first one this rule concerns each synthon, resulting in more diverse compounds.

Examples of analogs with different similarities to the initial molecules are given in **Figure 8 (III)**. As one can see, compounds with Tanimoto coefficient less than 0.5 are still quite similar to Molecule 1. Their distinctive feature is isomeric rearrangements in the position of substituents in the pyridine ring. Analogs with higher similarity mostly have pyridine substituted in the same positions as Molecule 1, which should increase not only structural but also shape similarity. Depending on the task in mind, the user can generate only the closest analogs with the default Synthl-Enumeration setup or also more diverse compounds by activating additional synthons selection with user-defined Tanimoto similarity threshold. This together with the ability to select reactions for bond disconnection/reassembling and BBs, provide a wide range of freedom for users.

## CONCLUSIONS

In this work, a new open-source toolkit for library design, called Synthons Interpreter or Synthl, was developed. It connects the building blocks (BBs) and fragments, derived from the pseudo-retrosynthetic fragmentation of larger compounds, via synthons-based representation. It is based on 38 reaction rules for bond disconnection. Their application results in a set of synthons that thanks to the presence of the special labels can be traced back to around 150 types of BBs. A herein-developed system of labels encodes the position and chemical nature of the reactive centers while preserving structure validity, allowing to treat synthons as actual compounds. Such an approach not only enables the design of synthetically accessible libraries but also facilitates BBs analysis in the medicinal chemistry context.

Here, Synthl was tested on the Enamine in-stock BB library for reagent classification, filtration and scaffold analysis. The list of recently approved drugs was used for compound fragmentation. The synthetic pathways

for those compounds reported in the literature were compared to Synthl results, demonstrating its accuracy in almost all cases, except heterocyclization steps, that have not been implemented yet. The analogs libraries were also generated for some of the drugs. The distinctive feature of Synthl library design is its strong dependence on the available BBs. Synthons-based library design allows generating collections of synthesizable compounds, that are structurally similar to the initial molecule and yet diverse with respect to each other.

## Supporting Information

Supporting information is available at <https://github.com/Laboratoire-de-Chemoinformatique/Synthl>.

## REFERENCES

1. Zhou, J. Z. Chemoinformatics and Library Design. In *Chemical Library Design*, Zhou, J. Z., Ed.; Humana Press: Totowa, NJ, 2011, pp 27-52.
2. Lenci, E.; Trabocchi, A., Smart Design of Small-Molecule Libraries: When Organic Synthesis Meets Cheminformatics. *ChemBioChem* **2019**, 20, 1115-1123.
3. Jamois, E. A., Reagent-based and product-based computational approaches in library design. *Curr. Opin. Chem. Biol.* **2003**, 7, 326-330.
4. Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M., On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, 3, 1503-1507.
5. Fechner, U.; Schneider, G., Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *J. Chem. Inf. Model.* **2006**, 46, 699-707.
6. Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M., RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 511-522.
7. ChemAxon, Ltd Budapest, Hungary, <http://www.chemaxon.com>.
8. OpenEye Scientific Software, Santa Fe, NM. <http://www.eyesopen.com>.
9. Landrum, G., RDKit: Open-Source Cheminformatics Software. <http://www.rdkit.org>.
10. Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E., AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **2020**, 12, 70.



11. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A., Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, 4, 522-532.
12. Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H., Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Organic Process Research & Development* **2015**, 19, 357-368.
13. Roughley, S. D.; Jordan, A. M., The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, 54, 3451-3479.
14. Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W., The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **2016**, 45, 546-576.
15. Cooper, T. W. J.; Campbell, I. B.; Macdonald, S. J. F., Factors Determining the Selection of Organic Reactions by Medicinal Chemists and the Use of These Reactions in Arrays (Small Focused Libraries). *Angew. Chem. Int. Ed.* **2010**, 49, 8082-8091.
16. Brown, D. G.; Boström, J., Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, 59, 4443-4458.
17. Tomberg, A.; Boström, J., Can easy chemistry produce complex, diverse, and novel molecules? *Drug Discov. Today* **2020**, 25, 2174-2181.
18. Hartenfeller, M.; Renner, S.; Jacoby, E., Reaction-Driven De Novo Design: a Keystone for Automated Design of Target Family-Oriented Libraries. *De novo Molecular Design* **2013**, 245-266.
19. Yasri, A.; Berthelot, D.; Gijzen, H.; Thielemans, T.; Marichal, P.; Engels, M.; Hoflack, J., REALISIS: A Medicinal Chemistry-Oriented Reagent Selection, Library Design, and Profiling Platform. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2199-2206.
20. Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J., SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, 46, 2765-2773.
21. 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 Combinatorial Science* **2012**, 14, 579-589.
22. Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G., DOGS: Reaction-Driven de novo Design of Bioactive Compounds. *PLOS Computational Biology* **2012**, 8, e1002380.
23. Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B., AllChem: generating and searching 1020 synthetically accessible structures. *J. Comput. Aided Mol. Des.* **2007**, 21, 341-350.
24. Bemis, G. W.; Murcko, M. A., The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, 39, 2887-2893.
25. Langdon, S. R.; Brown, N.; Blagg, J., Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, 51, 2174-2185.
26. Ruiz-Castillo, P.; Buchwald, S. L., Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chem. Rev.* **2016**, 116, 12564-12649.
27. Evans, G.; Blanchard, N.; Toumi, M., Copper-Mediated Coupling Reactions and Their Applications in Natural Products and Designed Biomolecules Synthesis. *Chem. Rev.* **2008**, 108, 3054-3131.
28. Korch, K. M.; Watson, D. A., Cross-Coupling of Heteroatomic Electrophiles. *Chem. Rev.* **2019**, 119, 8192-8228.
29. West, M. J.; Fyfe, J. W. B.; Vantourout, J. C.; Watson, A. J. B., Mechanistic Development and Recent Applications of the Chan–Lam Amination. *Chem. Rev.* **2019**, 119, 12491-12523.
30. Hughes, D.; Wheeler, P.; Ene, D., Olefin Metathesis in Drug Discovery and Development—Examples from Recent Patent Literature. *Organic Process Research & Development* **2017**, 21, 1938-1962.
31. Korotchenko, V. N.; Nenajdenko, V. G.; Balenkova, E. S.; Shastin, A. V., Olefination of carbonyl compounds: modern and classical methods. *Russian Chemical Reviews* **2004**, 73, 957-989.
32. Zajc, B.; Kumar, R., Synthesis of Fluoroolefins via Julia-Kocienski Olefination. *Synthesis* **2010**, 2010, 1822-1836.
33. Caro-Diaz, E. J. E.; Urbano, M.; Buzard, D. J.; Jones, R. M., C–H activation reactions as useful tools for medicinal chemists. *Bioorg Med Chem Lett* **2016**, 26, 5378-5383.
34. Bogolubsky, A. V.; Moroz, Y. S.; Mykhailiuk, P. K.; Pipko, S. E.; Konovets, A. I.; Sadkova, I. V.; Tolmachev, A., Sulfonyl Fluorides as Alternative to Sulfonyl Chlorides in Parallel Synthesis of Aliphatic

- Sulfonamides. *ACS Combinatorial Science* **2014**, 16, 192-197.
35. Kuttruff, C. A.; Haile, M.; Kraml, J.; Tautermann, C. S., Late-Stage Functionalization of Drug-Like Molecules Using Diversinates. *ChemMedChem* **2018**, 13, 983-987.
36. Proctor, R. S. J.; Phipps, R. J., Recent Advances in Minisci-Type Reactions. *Angew. Chem. Int. Ed.* **2019**, 58, 13666-13699.
37. Smith, J. M.; Harwood, S. J.; Baran, P. S., Radical Retrosynthesis. *Acc. Chem. Res.* **2018**, 51, 1807-1817.
38. Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A., Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **2018**, 10, 383-394.
39. Pennington, L. D.; Aquila, B. M.; Choi, Y.; Valiulin, R. A.; Muegge, I., Positional Analogue Scanning: An Effective Strategy for Multiparameter Optimization in Drug Design. *J. Med. Chem.* **2020**, 63, 8956-8976.
40. Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P., Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015**, 20, 11-17.
41. Goodman, J., Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, 49, 2897-2898.
42. Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; American Chemical Society: 2014; Vol. 1164, Chapter 8, pp 127-148.
43. Moo, Y. N., Ryune; LEE, Dae, Won; LEE, Ju, Young; KIM, Hui, Ho; LEE, Dong, Ho, , Method for preparation of carbamic acid (r)-1-aryl-2-tetrazolyl-ethyl ester. *World Intellectual Property Organization* **2010**, Patent number: WO2010/150946; A1.
44. Shivanyuk, A.; Ryabukhin, S. V.; Bogolubsky, A. V.; Tolmachev, A., Enamine REAL database: making chemical diversity real. *Chimica Oggi-Chemistry Today* **2007**, 58-59.
45. Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, 23, 101681.
46. Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J., The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016**, 56, 1253-1266.
47. Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H., Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, 49, 270-279.