Exploration of the Chemical Space of DNA-encoded Libraries

Regina Pikalyova, Yuliana Zabolotna, Dmitriy Volochnyuk, Dragos Horvath, Gilles Marcou, Alexandre Varnek

molecular informatics

Exploration of the chemical space of DNA-encoded libraries

Regina Pikalyova[a], Yuliana Zabolotna[a], Dmitriy M. Volochnyuk[c][d], Dragos Horvath[a], Gilles Marcou[a], Alexandre Varnek*[a][b]

Abstract: DNA-Encoded Library (DEL) technology has emerged as an alternative method for bioactive molecules discovery in medicinal chemistry. It enables the simple synthesis and screening of compound libraries of enormous size. Even though it gains more and more popularity each day, there are almost no reports of chemoinformatics analysis of DEL chemical space. Therefore, in this project, we aimed to generate and analyze the ultra-large chemical space of DEL. Around 2500 DELs were designed using commercially available BBs resulting in 2,5B DEL compounds that were compared to biologically relevant compounds from ChEMBL using Generative Topographic Mapping. This allowed to choose several optimal DELs covering the chemical space of ChEMBL to the highest extent and thus containing the maximum possible percentage of biologically relevant chemotypes. Different combinations of DELs were also analyzed to identify a set of mutually complementary libraries allowing to attain even higher coverage of ChEMBL than it is possible with one single DEL.

Keywords: DNA-encoded libraries, libraries design and comparison, GTM, drug design, hit identification

1 Introduction

Identifying compounds that bind to a biomacromolecule and show a desired therapeutic effect is a fundamental step in any drug discovery process. The most common method to find such molecules is high throughput screening (HTS)[1]. Since its emergence in the 1990s, HTS has delivered numerous lead molecules for drug development[2]. Nevertheless, this technology has several limitations, such as expensive robotic equipment and compound libraries, that are available mostly to large pharmaceutical companies[3]. The number of compounds that can be screened in one HTS campaign is usually limited to a million[4], while the chemical space of synthetically accessible molecules is far larger[5].

DNA-encoded library (DEL) technology has partially solved these problems[6]. It consists of the creation of libraries of DNA-encoded compounds using water-based combinatorial chemistry and their screening against soluble target proteins using binding affinity selection[7]. DNA-encoded compounds are molecules labeled with single- or double-stranded DNA. The latter plays a role of a “barcode” that encodes information about the building blocks (BBs) from which the compounds were synthesized. This DNA barcode allows to quickly identify successful ligands bound to the protein after affinity selection. The creation and screening of DELs offer many advantages compared to the conventional HTS approach. First of all, they are usually synthesized using a combinatorial split-and-pool approach[8] and thus allow to produce chemically versatile libraries of enormous size[9]. DEL compounds are screened all at once in a single vessel in contrast to individual compound screening in HTS[7]. Simple experimental setup of affinity selection accessible both in industry and university laboratories allows cheap and fast hits identification[10]. Many successful stories of employing this technology were published, including DEL-derived hits that progressed to the clinic[8].

However, up to this point, most efforts were focused on the analysis of the libraries of BBs or identified active compounds[3]. Authors were less keen to explore the entire chemical space covered by DELs because it is extremely vast. To our best knowledge, only one paper reported the analysis of DEL space using Reduced Complexity Molecular Frameworks (RCMF) methodology[11]. However, in that work, the analysis was limited to only four DELs (>5 × 10^8 compounds). Since DEL technology is actively being developed and new methodologies for DEL synthesis were being elaborated, the aforementioned pioneering work no longer reflects the status quo.

This work is focused on the generation of possible DELs from commercially available BBs using a tool for DELs generation called eDesigner[12]. Since screening thousands of DELs containing billions of compounds is unfeasible, we suggest choosing the so-called “golden” DEL(s) that covers the chemical space of biologically tested compounds to the highest extent. Such a library would have high structural diversity and contain the majority of biologically relevant chemotypes, which is critical for the success of the primary screening against novel biological targets. It was identified by comparing the

[a] *University of Strasbourg, Laboratory of chemoinformatics*
4, rue B. Pascal, Strasbourg 67081, France
**e-mail: varnek@unistra.fr, phone/fax: +33 368851560*

[b] *Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University*
Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan

[c] *Institute of Organic Chemistry, National Academy of Sciences of Ukraine*
Murmanska Street 5, Kyiv 02660, Ukraine

[d] *Enamine Ltd.*
78 Chervonotkatska str., 02660 Kyiv, Ukraine

Supporting Information for this article is available on the WWW under www.molinf.com

generated DEL space to the chemical space of biologically relevant ChEMBL[13] compounds using Generative Topographic Mapping (GTM) – an efficient dimensionality reduction method[14]. GTM has proved to be a powerful tool for "Big Data" analysis and visualization (up to 1B compounds)[15]. Notably, the prior development of quantitatively validated, polypharmacologically competent Universal Maps (uMaps) allowed us to propose a chemically meaningful representation of the to-date explored drug-like chemical space[16]. Only one of the several uMaps (uMap1, see the corresponding article) was used in this study for simplicity, but the study could be extended to consensus mapping on several uMaps.

2 Methods

2.1 General workflow

The workflow consists of seven parts, as shown in Figure 1. First, DEL-compatible chemical building blocks (BBs) were selected from the eMolecules and Enamine in-stock BB libraries described in the Data section. It was done on the basis of the Goldberg rule of two (Ro2)[17] and eDesigner built-in filters for selecting DNA-compatible BBs. Using these BBs, thousands of DELs were designed and generated with the help of eDesigner. The size of each DEL varied from 1M to 1B but for easier and quicker analysis, only a representative subset of 1M compounds per DEL was enumerated using the random sampling approach. In the third step, generated compounds were standardized according to the protocol explained in the Data section. ISIDA descriptors[18] were used to represent molecular structures in a machine-readable form of numerical N-dimensional vectors. They were then projected onto uMap1. Comparative landscapes were created and visualized to compare DEL compounds to biologically relevant molecules from the ChEMBL database. Then a so-called "golden" DEL that provides the highest coverage of ChEMBL chemical space was identified using responsibility patterns (RPs)[19]. To achieve even better coverage, complementary DELs were added to the "golden" one to give a "platinum" pool of DELs.

2.2 Selection of building blocks

Before DEL design and generation, input BBs were filtered according to Ro2 with the help of SynthI[20]. Ro2 is a guideline to choose high-quality BBs that can give access to drug-like molecules[17]. According to it, BBs should contribute to the final molecule only structural fragments that satisfy the following rules: MW < 200 Da, clogP < 2, number of H-bond donors ≤ 2, and number of H-bond acceptors ≤ 4. This filtration allows to limit the size of DEL compounds shifting corresponding libraries towards drug-like subspace of the chemical space. In addition to physicochemical properties, eDesigner built-in DNA-compatibility filters were also applied. The selection of building blocks by eDesigner is made by excluding compounds with unwanted functionalities that can lead to the reaction with water such as imines, benzyl halides, etc.

2.3 DEL generation with eDesigner

For the generation of chemical space of DELs, the eDesigner[12] tool was used. At first, based on the list of the most efficient DNA-compatible reactions encoded in the tool (see Supporting Information of respective article[12]) and a user-provided list of BBs, it generates a special set of instructions for DEL compound enumeration called libDESIGNs. Each libDESIGN contains information about the starting headpiece (the whole DNA part for computational convenience is formally represented as a ^{13}C atom), the reaction types, and BBs which will be used in them, as well as deprotection reactions for the final stage of DEL generation. There are also several restrictions that can be applied to control some of the properties of the resulting DEL. They include, for example, the maximum and the median value of heavy atom count in the generated molecules, minimum library size, etc. Once the libDESIGNs are created, the representative DELs subsets of the selected size can be enumerated by the LillyMol tool[21]. An example of such enumeration is shown in Figure 2. The isotopic mark on the carbon atom specifies the place of attachment of the DNA tag. For clarity reasons, before physicochemical properties calculation and GTM analysis, the ^{13}C atom is removed, therewith obtaining the compound that would have been resynthesized off-DNA for validation in case of being selected during a real screening campaign.

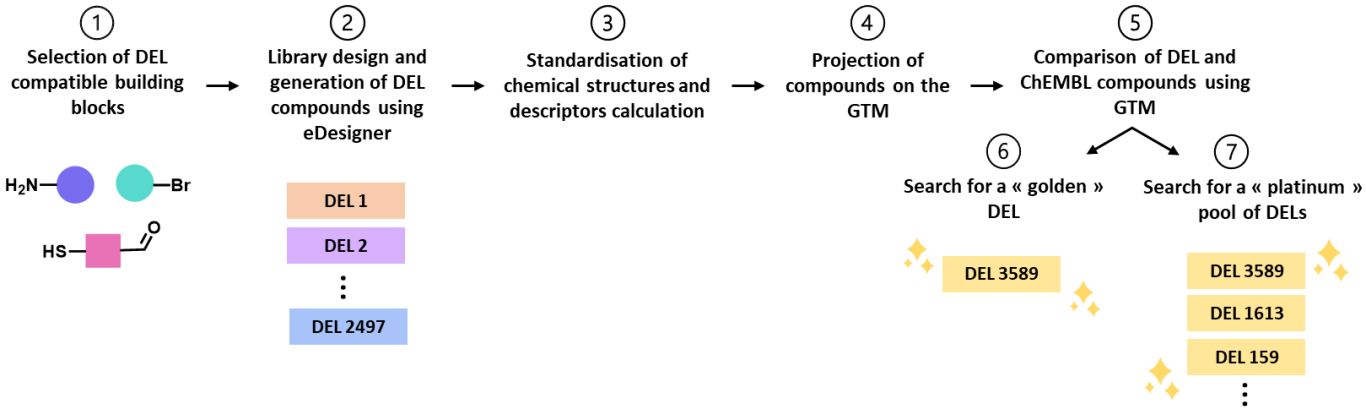

Figure 1. Workflow of the project. The rectangles represent separate DELs.

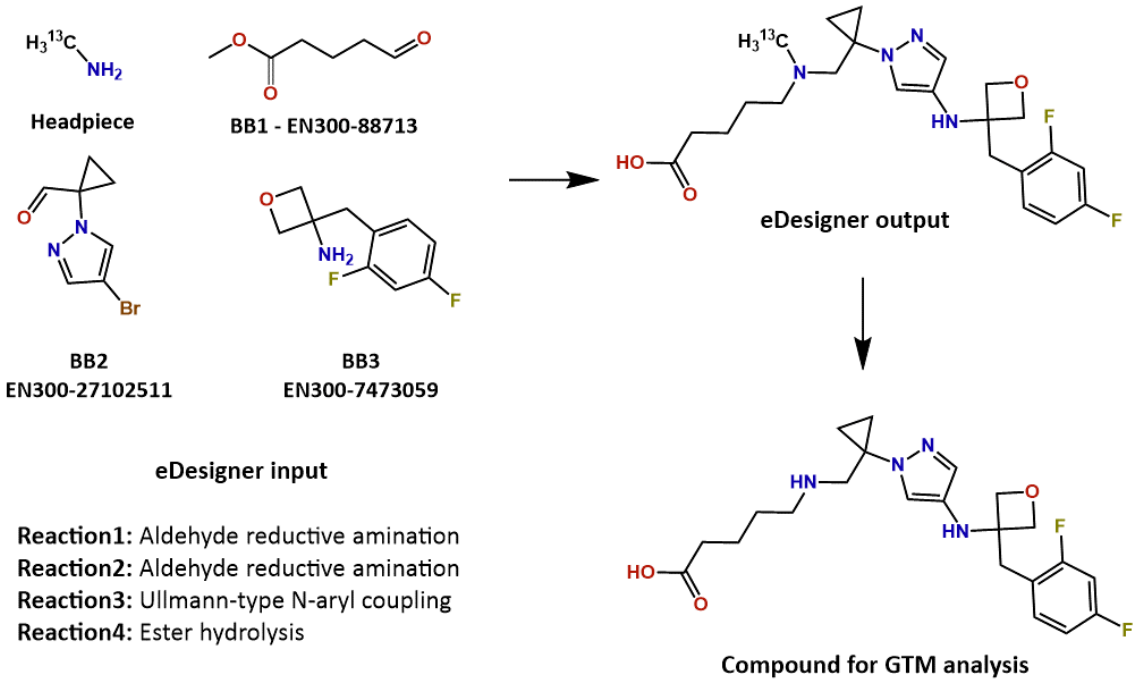

Figure 2. Example of DEL compound generation by eDesigner. The user should provide the headpiece and the list of BBs; an appropriate list of reactions will be selected automatically by eDesigner, and respective compounds will be generated. The isotopic mark is placed by eDesigner in order to know the position of DNA attachment and is removed prior to GTM analysis and physicochemical properties calculation.

2.4 Generative Topographic Mapping (GTM)

In the chemical space molecules are represented as data points, with their position being defined by a vector of numerical values called descriptors. The main idea of GTM[14] consists in inserting a flexible hypersurface called manifold into the high dimensional descriptor space with a subsequent projection of these data points into a 2D latent space grid.

The manifold is defined by a grid of Radial Basis Functions (RBFs, represented by Gaussian functions). It generates a probability distribution and is fitted to maximize the likelihood of the training set. The probability distribution generated by the GTM is evaluated over another grid of predefined locations, termed nodes. The number of RBFs is the key user-defined operational parameters; the number of nodes controls the map's resolution: it impacts the rendering but not the model itself. The GTM algorithm “bends” the manifold to pass through the densest areas of the data cloud formed by the points representing molecules of the input dataset. Then, the molecules are projected from the high-dimensional space onto the 2D map by associating each molecule to the several closest grid nodes. The degrees of association of each molecule to each node of the grid are called “responsibilities”. The responsibility of a node for a compound is the contribution of this node to the likelihood of this compound. Therefore responsibilities are real number vectors summing up to 1 over all nodes. Finally, the manifold is flattened out to obtain a 2D representation of the map with compounds projected onto it.

Based on the responsibility vectors, different types of landscapes can be created, where each node is colored using the weighted average of the properties of the compounds projected there. Properties assigned to each node are calculated as a weighted average of the properties of all residents, where weights are compound responsibilities to reside in this node. Depending on the information used for its coloration, there are two types of landscapes: class and property. The class landscape is used to analyze the distribution of the molecules of two classes in the chemical space. In this work, the class landscapes are used to visualize and analyze the distribution of the molecules of two classes – DEL (library 1) and ChEMBL (library 2) compounds. Property landscapes represent the distribution of molecular property or activity values. Using these landscapes, GTM can be applied for chemical space analysis, library comparison, or even virtual screening[22].

2.5 Universal GTM

The concept of Universal GTM (UGTM) was introduced by Sidorov et al.[23] and further developed by Casciuc et al.[16] as a general-purpose map that can accommodate ligands of diverse biological targets on the same GTM manifold. A genetic algorithm was used to choose the best descriptors set and GTM operational parameters (number of nodes and RBFs, manifold flexibility controls, etc.) so as to maximize the mean predictive performance over hundreds of biological activities from ChEMBL. The resulting best uMap1 allowed to separate molecules by their activity class (active/inactive) against 618 (later extended to 749) biological targets, which makes it “polypharmacologically competent”. This map was built based on ISIDA atom sequence counts with a length of 2−3 atoms labeled by CVFF force field types and formal charge status[18]. The size of the map was chosen to be 41x41 nodes and the number of RBFs - 18x18.

Since the ChEMBL database is the most reliable source of the compounds with experimentally measured

biological activity[13], the universal maps trained on the ChEMBL data series are highly oriented towards biologically relevant compounds. Apart from predicting biological activity, these maps can also be used as frameworks for analyzing large chemical libraries in medicinal chemistry and drug design context. The uMap1 has been used in this project to compare biologically relevant compounds from ChEMBL with the DNA-encoded compounds. This choice was motivated by previous results in identifying biologically relevant molecules missing from the chemical market, as well as untested commercially available compounds when comparing ChEMBL and ZINC[15].

2.6 Responsibility patterns

As mentioned previously, compounds are mapped on the GTM with certain responsibilities - probabilities of these compounds to populate a specific node of the map. Since these values are real numbers, finding two molecules with identical responsibility vectors is highly improbable. This makes it challenging to identify structurally similar compounds by their responsibility vectors – they may be slightly different even for very similar compounds. To solve this problem, it was suggested by Klimenko et al.[19] to discretize the vector, with all responsibility values less than 0,01 being reassigned to zero and all others - to a number from 1 to 10. This discretized vector is referred to as Responsibility Pattern (RP) and is calculated for each compound according to the formula in Figure 3.

Molecules whose R vectors round up to the same RP are considered to be grouped in the same cell of the chemical space and thus to form a cluster of similar structures[22]. For example, in Figure 3, a GTM density landscape, featuring compound sets associated with two different RPs is shown. Colors encode the cumulative sum of responsibilities of all compounds residing in the particular node (grey regions are moderately populated, while colored ones contain a higher number of compounds). RP1 corresponds to the 221 indoles that contain additional amino and/or guanidino functional groups. These compounds occupy a small compact area of the chemical space distanced from the island of RP vector 2, populated by 173 naphthols, polyphenols, and their methyl ethers. In this work, RPs were used to compare each separate DEL to ChEMBL, i.e. to evaluate the proportion of ChEMBL RPs ("structural motifs") also covered by a given DEL.

2.7 ChEMBL coverage estimation

First, RPs for all compounds are calculated as described above. Then the pairwise overlap between each DEL and ChEMBL is determined by dividing the number of common RPs for both libraries by the total number of ChEMBL RPs:

$$ChEMBL\ RPs\ coverage\ \% = \frac{Number\ of\ ChEMBL\ RPs\ present\ in\ DEL}{Total\ number\ of\ ChEMBL\ RPs}$$

However, the analysis of the percentage of covered ChEMBL RPs does not consider the number of compounds corresponding to each RP, although different RPs can be populated differently – from 1 to ≈12 000 compounds. As a result, increasing RP coverage does not necessarily mean significantly increasing the compound coverage. Thus the ChEMBL RPs coverage (%), weighted by RP population (the number of ChEMBL compounds per RP), is also used:

$$Weighted\ ChEMBL\ RPs\ coverage\ \% = \frac{\sum Population\ of\ ChEMBL\ RPs\ present\ in\ DEL}{\sum Population\ of\ all\ ChEMBL\ RPs}$$

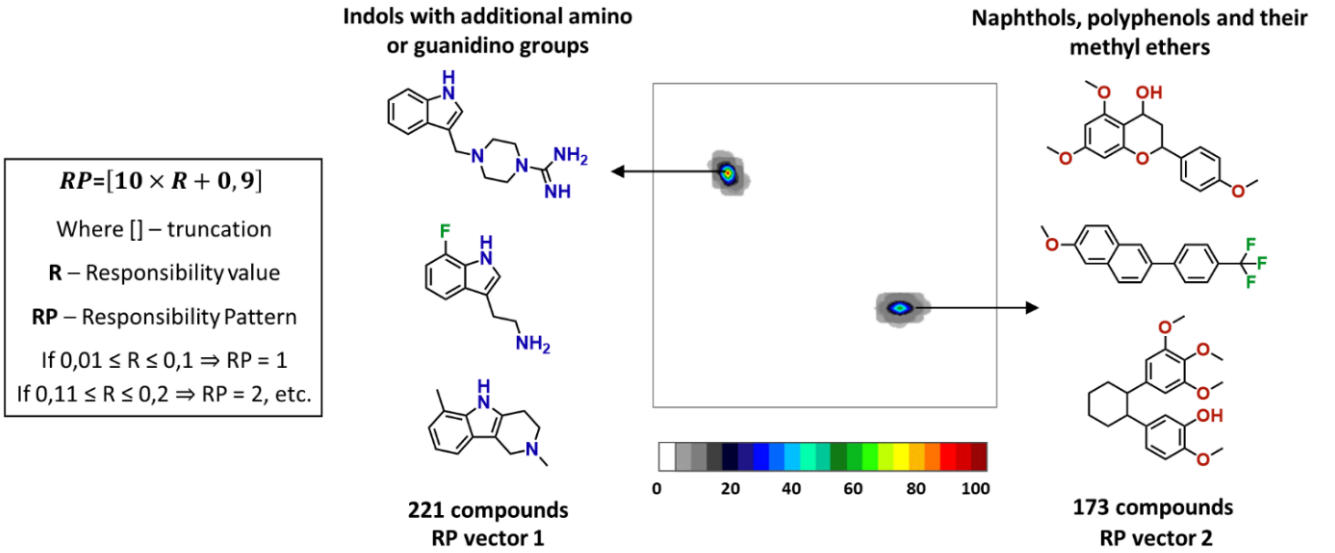

Figure 3. Left: formula for responsibility pattern (RP) calculation. Right: example of compounds sharing the same RPs and their position on the density landscape - a map colored by the local density of compounds. Highly populated zones are colored in red, underpopulated ones - in grey.

3 Data

3.1 Commercially available BBs

A set of 450K commercially available BBs was provided by eMolecules Inc.[24] They were complemented by an "orthogonal" (i.e. containing completely different BBs) dataset of 10K Enamine[25] in-stock BBs. Among them, only 79 141 BBs that satisfy Ro2 and eDesigner built-in DNA-compatibility filters were selected.

3.2 ChEMBL (biologically tested compounds)

ChEMBL is a database containing >2M diverse and biologically relevant compounds against >14K biological targets[13]. The major goal of this project was to find structurally diverse DELs suitable for primary screening. Since similar structures tend to have similar properties, finding a DEL containing compounds structurally similar to ChEMBL means finding a DEL that contains biologically relevant molecules. Such DEL will have a high potential to contain hit compounds. Hence, ChEMBL (version 28) was used as a reference library that guides our choice of the best DEL for primary screening. First, 2 086 898 molecules were downloaded from ChEMBL. After standardization, 1 853 565 unique compounds with known biological activities remained. The standardization of chemical structures was done using ChemAxon Standardizer[26] according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics in the University of Strasbourg[27]. It included dearomatization and final aromatization (heterocycles like pyridone are not aromatized), dealkalization, conversion to canonical SMILES, removal of salts and mixtures, neutralization of all species, except nitrogen(IV), generation of the major tautomer according to ChemAxon. After the standardization, the ISIDA fragment descriptors used to construct the first universal map (described in Experimental section 4) were calculated for all molecules. The same procedure was also applied to generated in this work DEL compounds.

4 Results and discussion

4.1 DNA-compatible BBs and reactions for DEL generation

The scope of synthetic procedures used in DEL chemistry is limited to high-yielding DEL-compatible reactions. Synthetic efforts to adapt reactions for use in DEL technology have been underway for several years, but the number of optimized for DEL chemistries is still rather restricted[28]. For example, only a few heterocyclisations optimized for DEL synthesis were described, such as benzimidazole, imidazolidinone, thiazole synthesis, and some others[29]. Nevertheless, even a few reactions can give rise to structurally diverse DELs if abundant building blocks (BBs) sets are employed for their generation.

In this work, 79 141 mono-, bi-, and trifunctional BBs were used for DEL generation. They were obtained by applying the Goldberg rule of two and built-in eDesigner DEL-compatibility filters to the combined in-stock library provided by eMolecules and Enamine. Prevalent monofunctional BB classes in the resulting dataset are secondary and primary amines, aryl halides, and carboxylic acids (Figure 4).

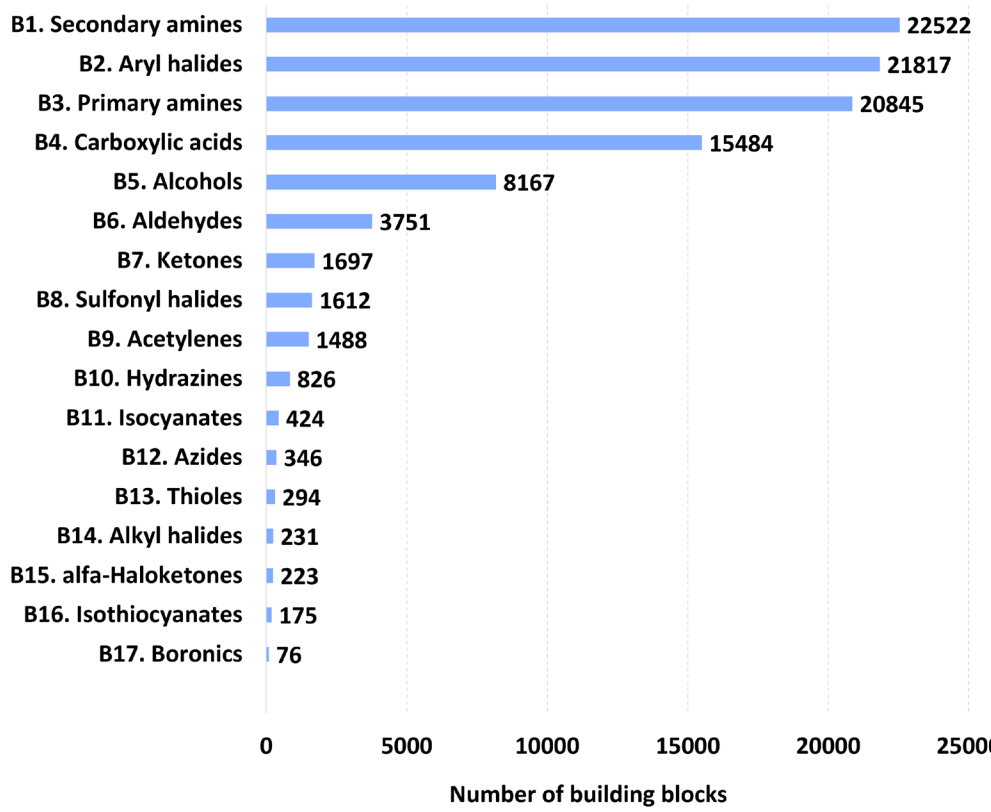

Figure 4. Monofunctional DNA-compatible commercially available BBs.

Due to their participation in common DNA-compatible combinatorial reactions (such as condensation of carboxylic acids with amines, aldehyde reductive amination, bromo-Sonogashira coupling, etc.), there is an active development of such BBs, making these four classes more structurally rich and widely available commercially. Note that in this work, all structures were stereochemistry-depleted (a unique skeleton graph is used to represent all stereoisomers). Therefore, the number of different BBs is higher.

In the case of bifunctional BBs (Figure 5), protected amino acids (AA) (such as amino esters, N-Boc-AA, N-Fmoc-AA, etc.) represent the most abundant class (3 796). The reason for such abundance is the popularity of peptide bond formation for DEL compounds' synthesis that requires this type of reagents. However, the number of actual AA fragments available from BBs with multiple protective groups is slightly smaller (2 885). It appears that the majority of AA fragments (2 173) occur in only one protected form, and 712 AA were found in the library more than once with different protecting groups. Figure 6 (I) shows an example of AAs that occur in the maximum number of protected variations in the BB library.

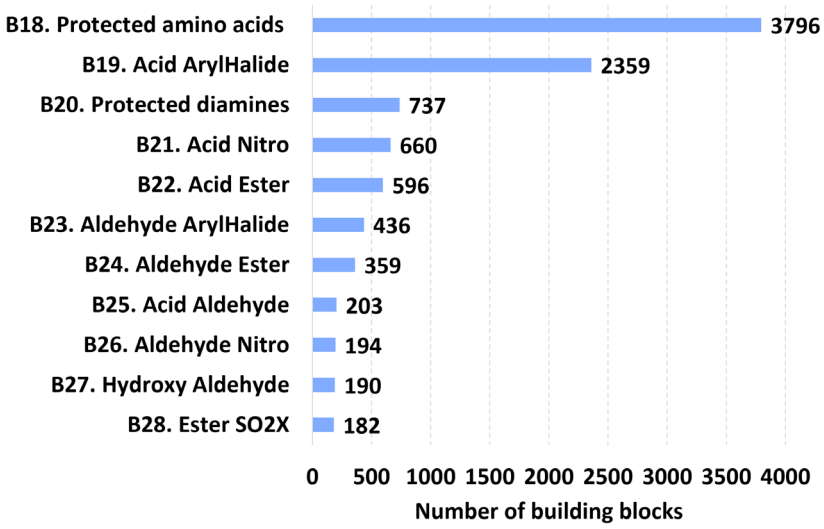

Figure 5. Bifunctional DNA-compatible commercially available BBs.

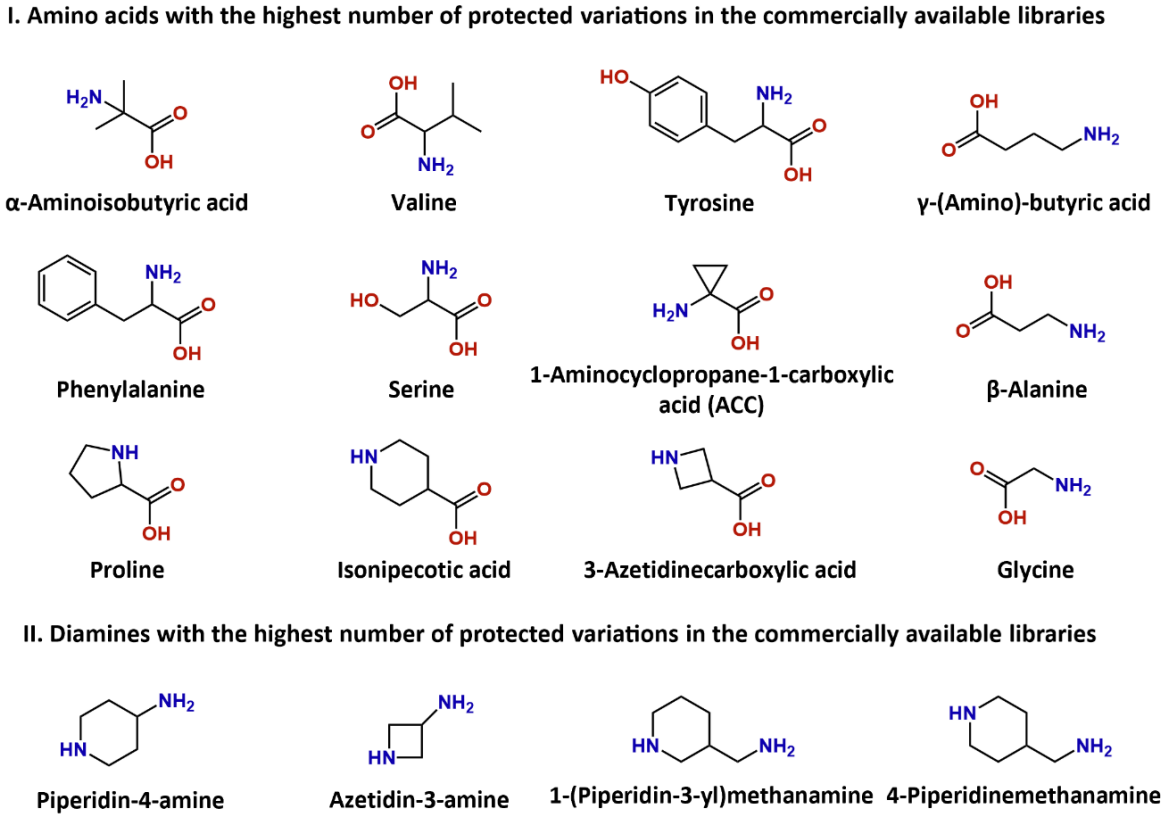

Figure 6. AA (I) and diamines (II), represented in the commercially available libraries of DNA-compatible BBs with the highest number of protected variations (N-Boc, N-Fmoc, various esters, etc.).

A similar tendency is also observed for protected diamines that occupy third place in the bar chart in Figure 5 after BBs containing both aryl halide and carboxylic acid functionality (2 359). A total of 737 protected diamines are equivalent to only 632 unique diamine fragments. Among them, 510 are represented by only one protected variant, while the other 122 occur in several differently protected copies. Four diamines, each occurring in the highest number of protected variations, are shown in Figure 6 (II). The number of trifunctional BBs is significantly lower than other reagents due to higher structural complexity (Figure 7). The most highly populated class of trifunctional BBs is haloaryl nitrocarboxylic acids containing 110 members. In DEL technology nitro group usually pose as a latent amino group that can be obtained upon reduction.

Using these BBs and user-defined library limitations in eDesigner, 2 497 DELs were designed. The maximal number of heavy atoms in DEL compounds was set to be 45, and at least half of all compounds in the library needed to have less than 35 non-hydrogen atoms. The frequency of the use of a particular reaction to generate all DELs is shown in Figure 8.

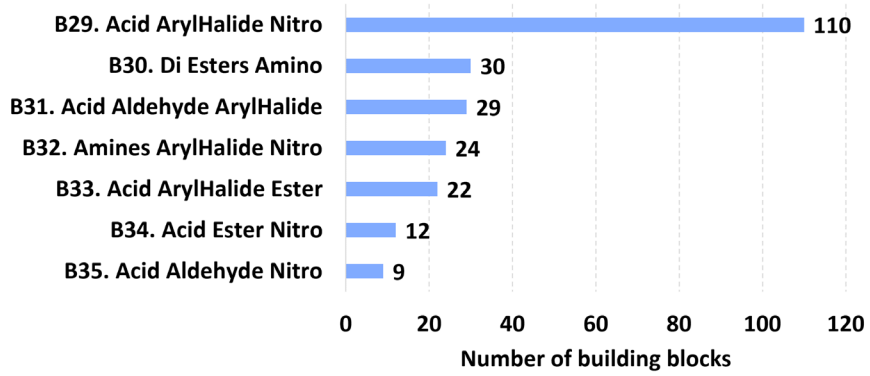

Figure 7. Trifunctional DNA-compatible commercially available BBs.

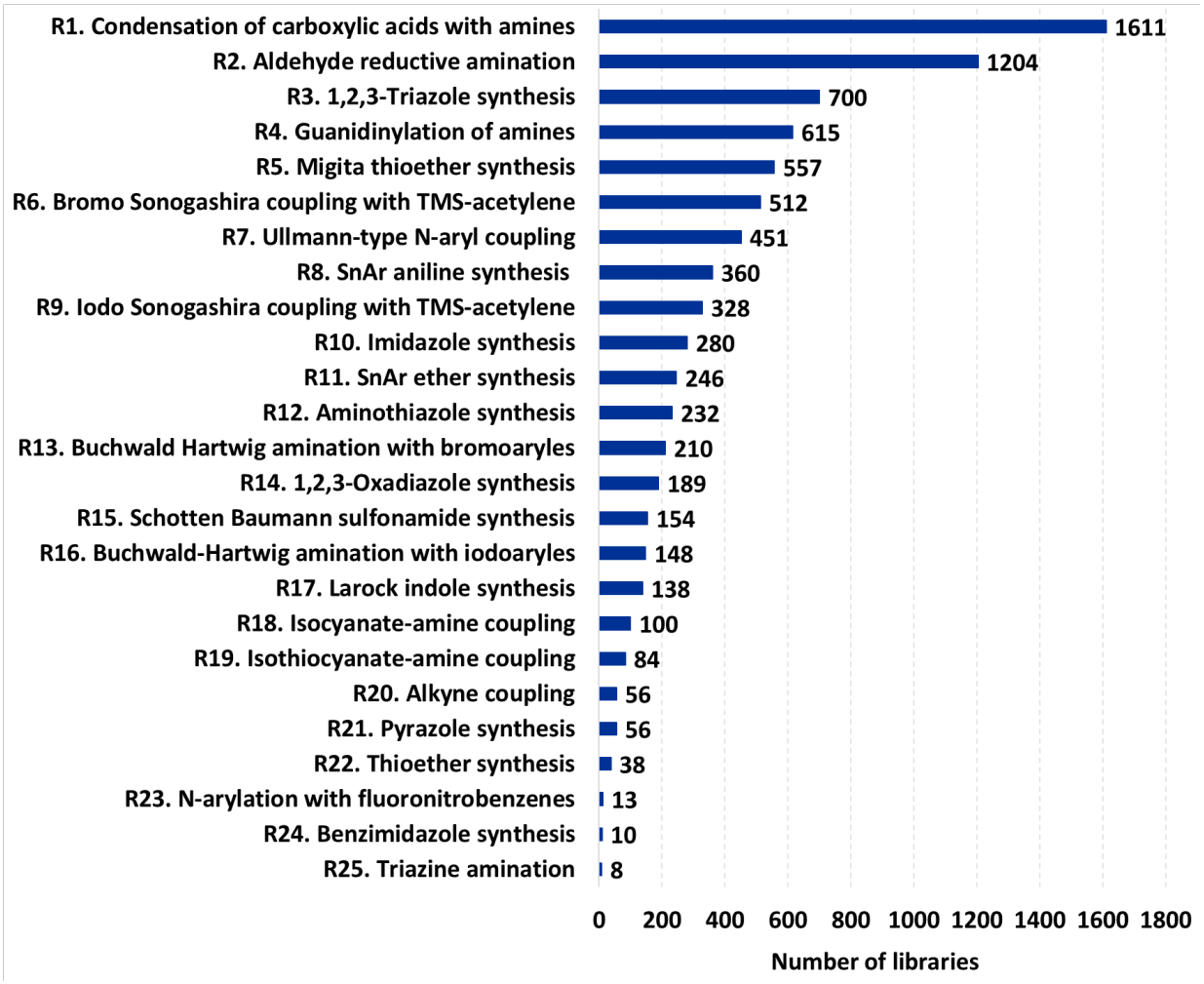

Figure 8. Frequency of the use of a particular reaction in DELs generation.

The most frequently used reactions, each being exploited in more than 500 libraries, were: condensation of carboxylic acids with amines (R1), aldehyde reductive amination (R2), 1,2,3-triazole synthesis (R3), guanidinylation of amines (R4), Migita thioether synthesis (R5), and bromo-Sonogashira coupling with TMS acetylene (R6). The high frequency of reaction usage is mainly caused by the prevalence of the respective BB classes in the input library (B1, B2, B3, B4 in Figure 4). Indeed, the amines are coupling partners in three reactions mentioned above (R1, R2, and R4), aryl halides - in two (R5 and R6), and carboxylic acids in R1.

Not all compounds were enumerated for every DEL, but random sets of 1M representative compounds were produced by eDesigner. In order to verify that such a library core is indeed representative, the whole library of 88M was enumerated for one of the DELs, and density landscapes were built for the whole library and 1M dataset on the same density scale. As one can see in Figure 9, each region of the map, occupied by the members of the whole library, also has representatives in the 1M randomly generated dataset – colored regions coincide on both maps, and only the density of residents differs. Therefore, 1M randomly enumerated compounds will be considered in this work as a sufficient representation of the whole DEL for GTM-based analysis.

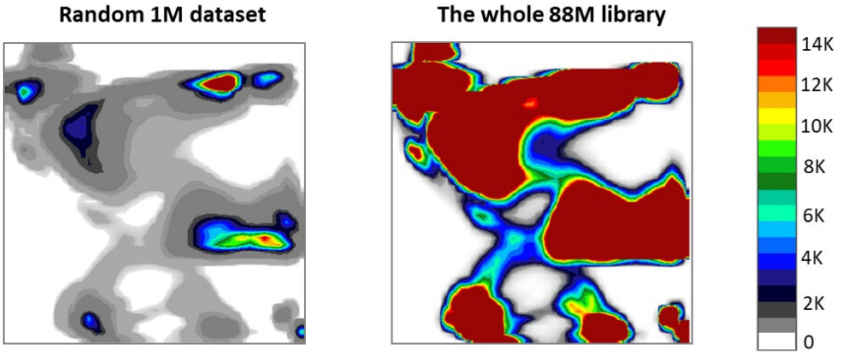

Figure 9. Comparison of the density distribution for the 1M randomly generated compounds and the whole DEL (88M). The color scale encodes the corresponding number of compounds residing in each colored node of the map.

4.2 Physicochemical properties of generated libraries

Out of a total of 2 497 generated DELs, 77 are produced by a single coupling reaction of 2 BBs (hence the label "2BB libraries"). The remaining 2 420 DELs are "3BB libraries". The physicochemical properties were calculated using RDKit[30]. Drug-like[31] (MW≤500; LogP≤5; the number of H-bond donors≤5; the number of H-bond acceptors≤10; ring counts≤10) and lead-like[32] (MW≤400; -3.5≤LogP≤4; the number of H-bond donors≤5; the number of H-bond acceptors≤8; ring counts≤4; rotatable bonds≤10) filters were applied. Figure 10 depicts how many of 2BB and 3BB libraries (in percentage) contain a specified portion of drug-like (Figure 10 (I)) and lead-like (Figure 10 (II)) compounds.

As expected, 2BB libraries contain smaller compounds, and thus the portion of drug- and lead-like compounds for them is higher than for 3BB DELs. For almost a half of 2BB libraries, all generated compounds fall into the category of drug-like, while in the case of 3BB DELs, only 2% of libraries are fully drug-like. However, the content of such compounds in 3BB libraries is still relatively high – the majority of DELs (68%) contain at least 50% of drug-like compounds. At the same time, the number of lead-like compounds is significantly lower for both categories of DELs. Almost a quarter of all 2BB libraries do not contain them, and another quarter is less than 50% lead-like. In the case of 3BB libraries, the lead-like compounds are almost entirely absent – 70% of DELs do not contain such molecules at all, and the remaining 30% of libraries have only up to 30% of lead-like molecules.

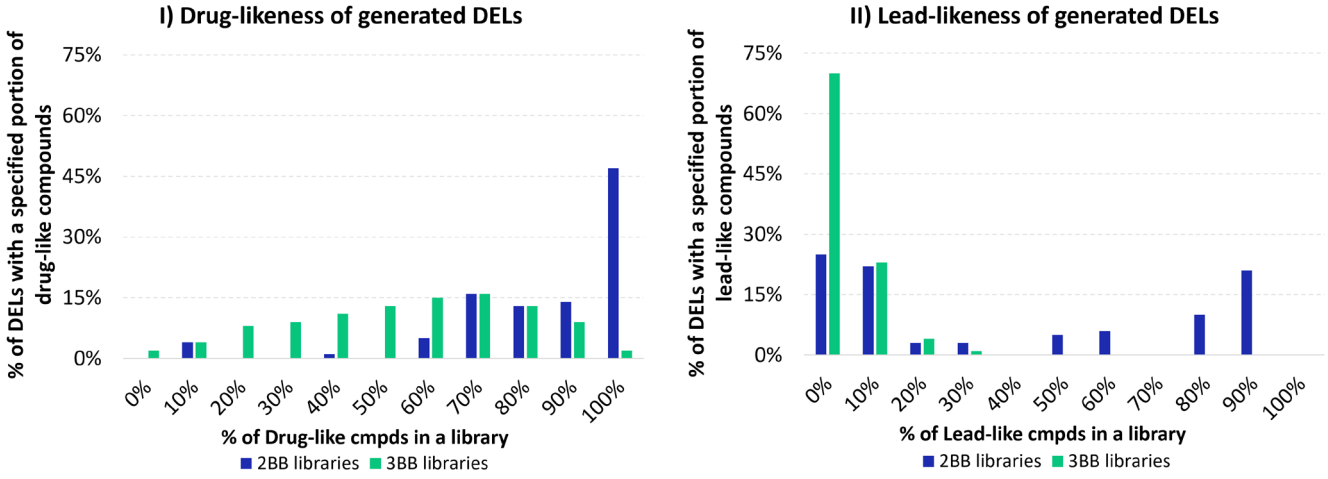

Figure 10. Comparison of (I) drug- and (II) lead-likeness of 2BB and 3BB libraries: percentage of 2BB and 3BB libraries having a particular portion of compounds satisfying respective filters is given.

4.3 Search for the "golden" DEL

The "golden" DEL can be defined as a library that is diverse enough to cover the highest possible proportion of biologically relevant compounds from ChEMBL. This coverage was calculated in terms of common responsibility patterns (RPs) explained in the Methods section. In Figure 11 (a) one can see the number of libraries with particular coverage of ChEMBL RPs. The majority of libraries cover 10-20% of ChEMBL chemical space in terms of unweighted RPs coverage score. 64 DELs showed the highest coverage of ChEMBL RPs – 30-33%. Figure 11 (b) depicts the coverage of the ChEMBL RPs weighted by the number of compounds that correspond to each RP. This time, 90 DELs showed high coverage of ChEMBL chemical space, ranging from 50 to 60%.

Figure 12 displays three comparative landscapes: DEL1857 with 13%, DEL167 with 27%, and DEL3589 with 57% coverage of ChEMBL (here, weighted coverage is considered). Dark grey zones are populated exclusively by ChEMBL molecules, while all other colors indicate areas also containing DEL compounds in a different ratio. Below each landscape, the IDs of reactions used for the corresponding library generation are given (see Figure 8 for reaction IDs). From the landscape of DEL1857, it is apparent that this library does not cover many areas of ChEMBL chemical space – there are few multi-colored spots on the landscape. It is an indicator that DEL1857 is not chemically diverse enough, and there are plenty of biologically relevant chemotypes absent from this library. DEL167, in its turn, allows achieving higher coverage of ChEMBL. DEL3589, on the other hand, is one of the leaders among all 2,5K DELs - multi-colored areas are not focused in one place of the map, but rather distributed on different islands that correspond to different chemotypes, and dark grey areas are less present.

There are around 90 libraries with similar chemical space coverage and diversity, but here, we will limit the discussion to the DEL3589 as an example of a "golden" DEL. The 84 M compounds of this DEL can be obtained by a succession of three reactions: two aldehyde reductive amination steps followed by Ullmann-type N-aryl coupling (see Figure 14, DEL3589). BBs used are 3 138 aldehydes, 275 bromoarylaldehydes, and 97 amines. As was discussed earlier, the latter is the class with the highest number of diverse BBs (Figure 4). Therefore, a random selection of BBs for DEL generation from such various and numerous collection results in higher coverage of ChEMBL chemical space.

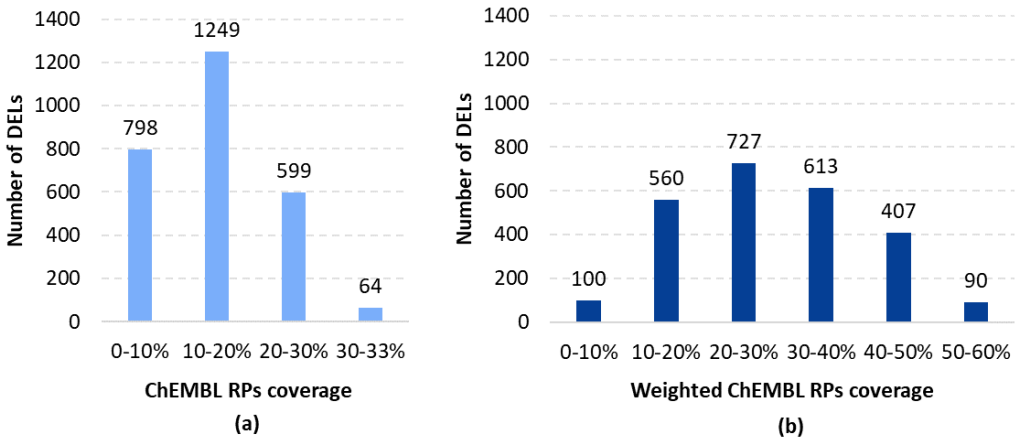

Figure 11. (a) Number of DELs with different coverage of ChEMBL responsibility patterns (RPs) (b) Number of DELs with different percentages of ChEMBL RPs coverage weighted by the RPs population (number of ChEMBL compounds per RP).

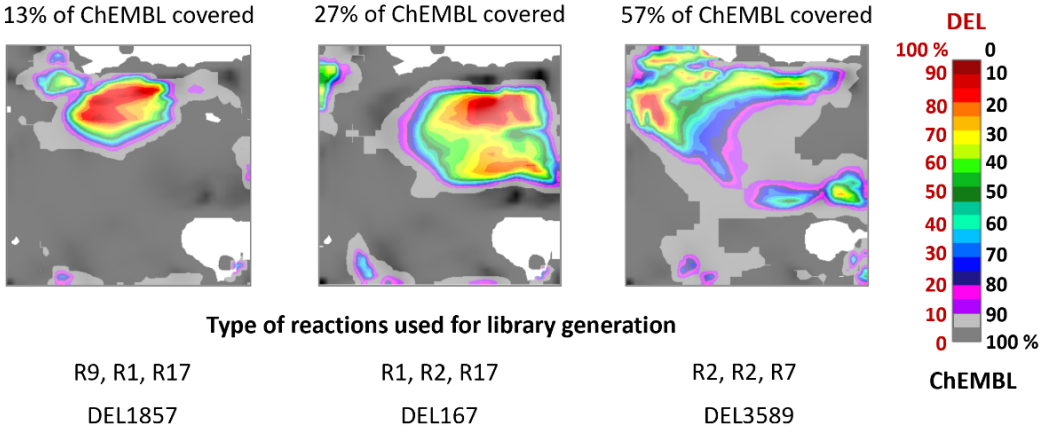

Figure 12. Class landscapes comparing a particular DEL with ChEMBL. From left to right: comparison of ChEMBL to DEL1857, DEL167, and DEL3589. Dark grey zones are populated exclusively by ChEMBL compounds, while all other colors indicate areas also containing DEL compounds in a different ratio. White regions correspond to the empty areas of the chemical space. Below each landscape, a library ID and IDs for corresponding reaction types are given.

4.4 Search for the "platinum" set of DELs

As shown on the class landscape for DEL3589 in Figure 12, there are still some dark-grey zones left that are not covered even by this "golden" DEL, which means there is space for improvement. To fill uncovered parts of the chemical space, the approach of library pools[33] was considered. According to it, several distinct DELs may be combined to create a more complex mixture, called "library pool", which can then be screened all at once. In order to obtain the highest coverage of ChEMBL, composing DELs for constructing such library pools should be complementary to each other, and each new DEL should cover previously unrepresented areas of the biologically relevant space.

The 90 DELs with the highest weighted coverage of ChEMBL RPs were chosen as possible "root" library. Each of these was then iteratively completed with up to 14 other libraries. Every complementary DEL was chosen in a way to cover the maximal portion of the ChEMBL chemical space that was not covered in the previous steps. Each time a complementary DEL was added to the pool, the weighted ChEMBL coverage was calculated. The line chart in Figure 13 was used to identify a pool of DELs that can enhance ChEMBL coverage to the highest possible extent. It shows how the weighted ChEMBL coverage increases over the addition of complementary libraries. According to this chart, after the fifth DEL, each complementary library provides less than 1% of additional weighted ChEMBL coverage – irrespectively of the chosen root DEL. Considering that the size of each DEL can vary from 1M to 1B compounds, adding a library of such large size to the pool only to increase ChEMBL coverage by 1% is not justified. Therefore, it is irrational to use a pool of DELs composed of more than five libraries.

If above-described DEL3589 is used as root DEL, the "platinum" pool of five DELs will be composed of such libraries: DEL3589, DEL1613, DEL159, DEL1161, and DEL845. Overall, they contain 776M compounds. Reactions used for the generation of these five DELs are shown in Figure 14: aldehyde reductive amination (R2), Ullmann type N-aryl coupling (R7), condensation of carboxylic acids with amines (R1), guanidinylation of amines (R4), and SnAr ether synthesis (R11). Almost all of them are among the most frequently used reactions for DEL generation (Figure 8) that employ BBs from highly represented classes (Figure 4). On the other hand, a pool of three DELs (DEL3589, DEL1613, DEL159) can be even more convenient since it contains fewer compounds (around 487M) and yet still allows to cover a large portion of ChEMBL (77%).

The physicochemical properties of the selected libraries were calculated and analyzed (Table 1). The proportion of drug-like and lead-like compounds varies for all DELs. The 2BB DEL159 shows the highest percentage of drug-like and lead-like molecules, 98% and 78%, respectively. This result is not surprising due to the lower molecular weight of compounds from 2BB libraries. Regarding 3BB libraries, it appears that the golden DEL3589 possesses higher drug-likeness (80% of such compounds) and lead-likeness (12% of such compounds) than the 3BB complementary DELs. Indeed, 52% of molecules from DEL1613 are drug-like while for DEL1161 the proportion of such compounds is only 30%. The portion of lead-like molecules for these libraries is negligible.

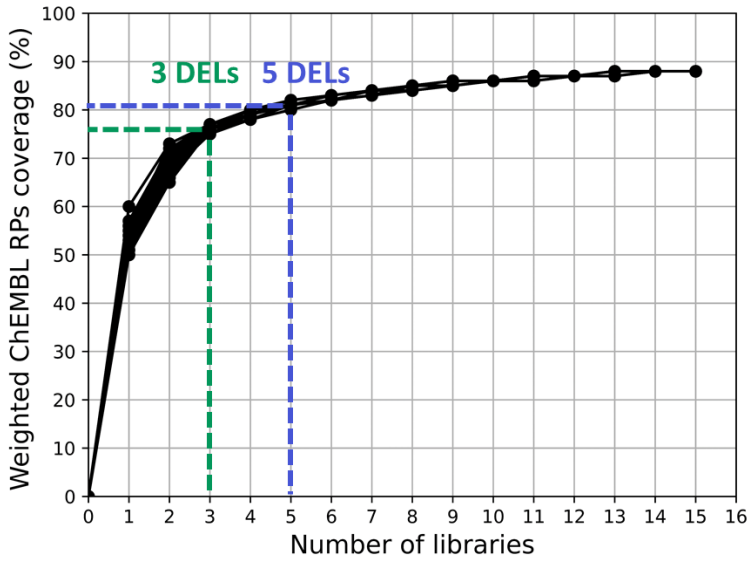

Figure 13. The percentage of the ChEMBL coverage, weighted by the number of compounds sharing common RPs, as a function of the number of libraries in the set. Green and blue dashed lines highlight the points for three and five DELs.

Table 1. The percentage of drug-like and lead-like compounds in the selected DELs that form "platinum" pools of three and five DELs. All DELs are 3BB libraries except DEL159 which is a 2BB library.

	% drug-like compounds	% lead-like compounds
DEL3589	80%	12%
DEL1613	52%	5%
DEL159	98%	78%
DEL1161	31%	1%
DEL845	71%	6%

To better illustrate how ChEMBL coverage increases when a pool of DELs is used instead of a single DEL, four comparative landscapes – featuring the "golden" DEL, the "platinum" pools of three and five DELs, and ≈2,5K DELs against ChEMBL were created (Figure 15). Structural analysis of underrepresented in DELs zones was carried out (Figure 16). The obtained landscapes show that as we go from one (Figure 15 (I)) to three DELs (Figure 15 (II)), the ChEMBL coverage increases drastically.

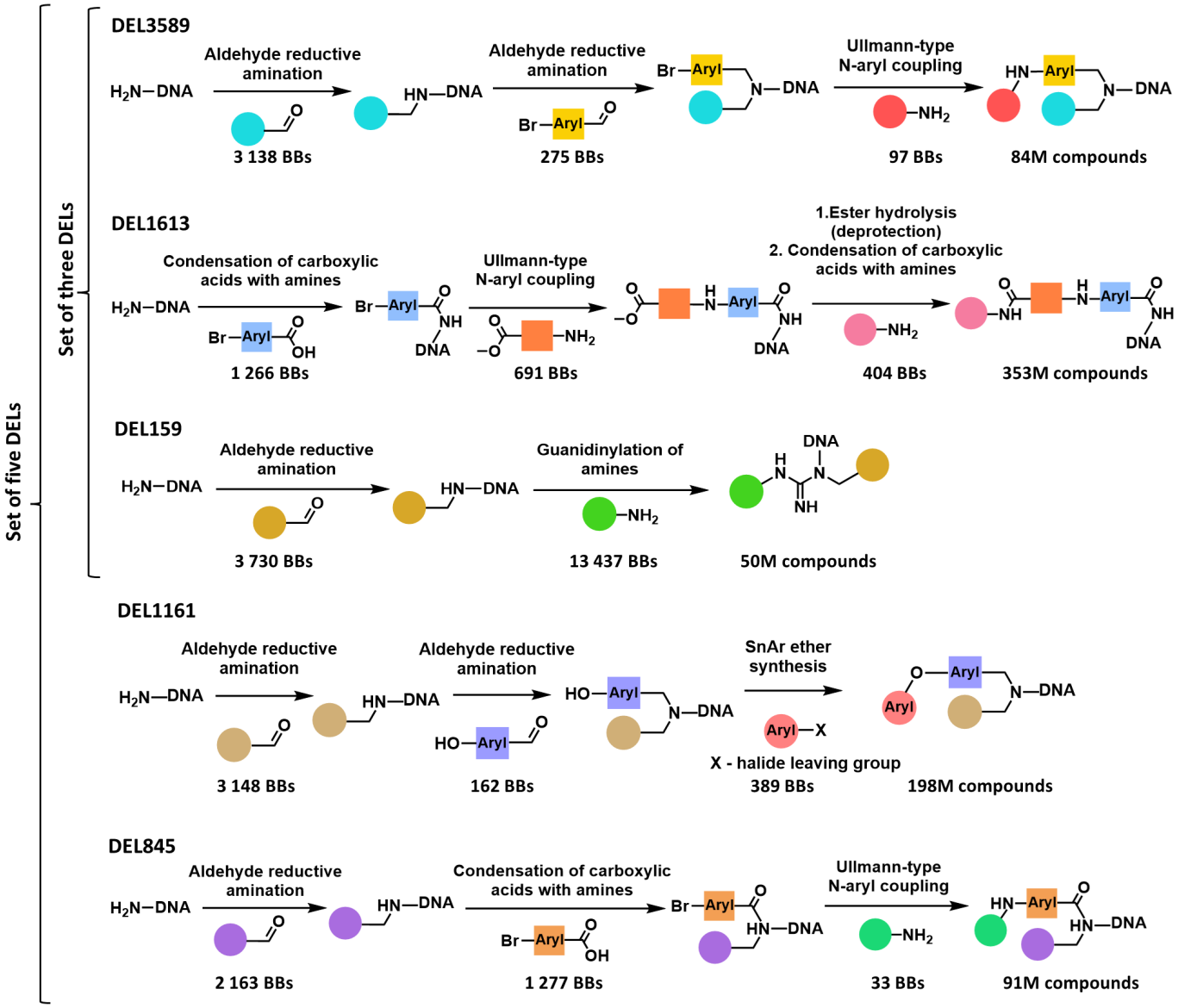

Figure 14. Reactions and BBs required for the synthesis of the "golden" DEL and libraries composing "platinum" pools.

On the landscape of the "platinum" pool of three DELs, the ChEMBL areas from A1 to A6 became a lot more populated. However, the addition of the following two libraries does not have the same impact. There are almost no new previously uncovered areas, only the increase in the population of previously occupied areas is observed (Figure 15 (III)). However, neither three nor five libraries succeeded in covering areas A7 and A8 completely. To see whether it is even possible to do so, a comparative landscape for all DELs versus ChEMBL was created (Figure 15 (IV)). It appears that neither of the DELs can cover these regions of the chemical space – areas A7 and A8 remained dark-grey. This result is not surprising because they contain natural products (NP) and NP-like compounds such as cardiac glycosides, steroids, and steroid-like compounds, saccharides, nucleotides, oligopeptides, coumarins, macrolides, chalcones, etc., which are indeed inaccessible by DEL technology as employed in this analysis.

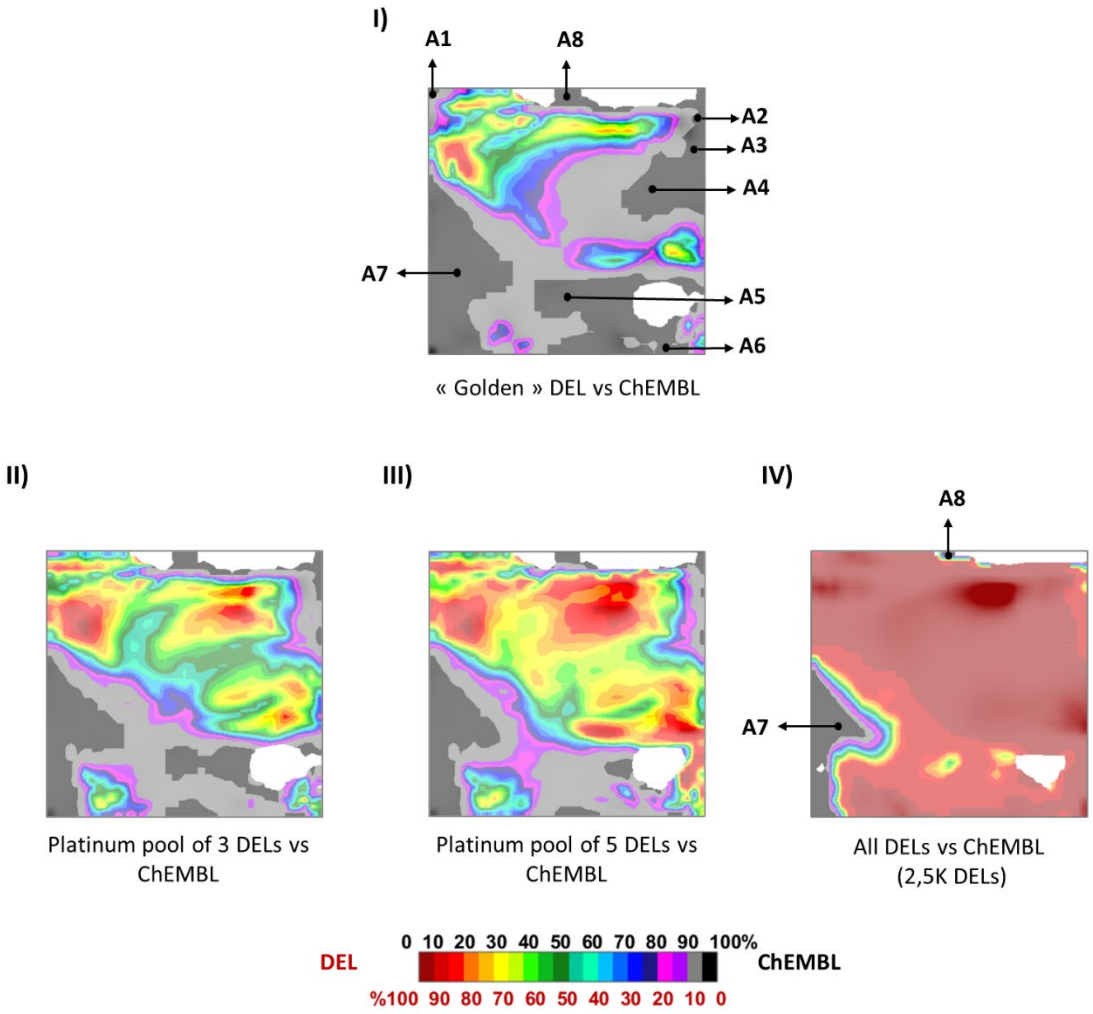

Figure 15. Comparison of ChEMBL and I) "golden" DEL, II) a pool of three DELs, III) a pool of five DELs, and IV) all 2,5K DELs. Multicolored zones are populated by both ChEMBL and DEL compounds, dark grey zones – only by ChEMBL compounds. White regions correspond to the empty areas of the chemical space. Examples of compounds populating highlighted areas A1-A8 are provided in Figure 16.

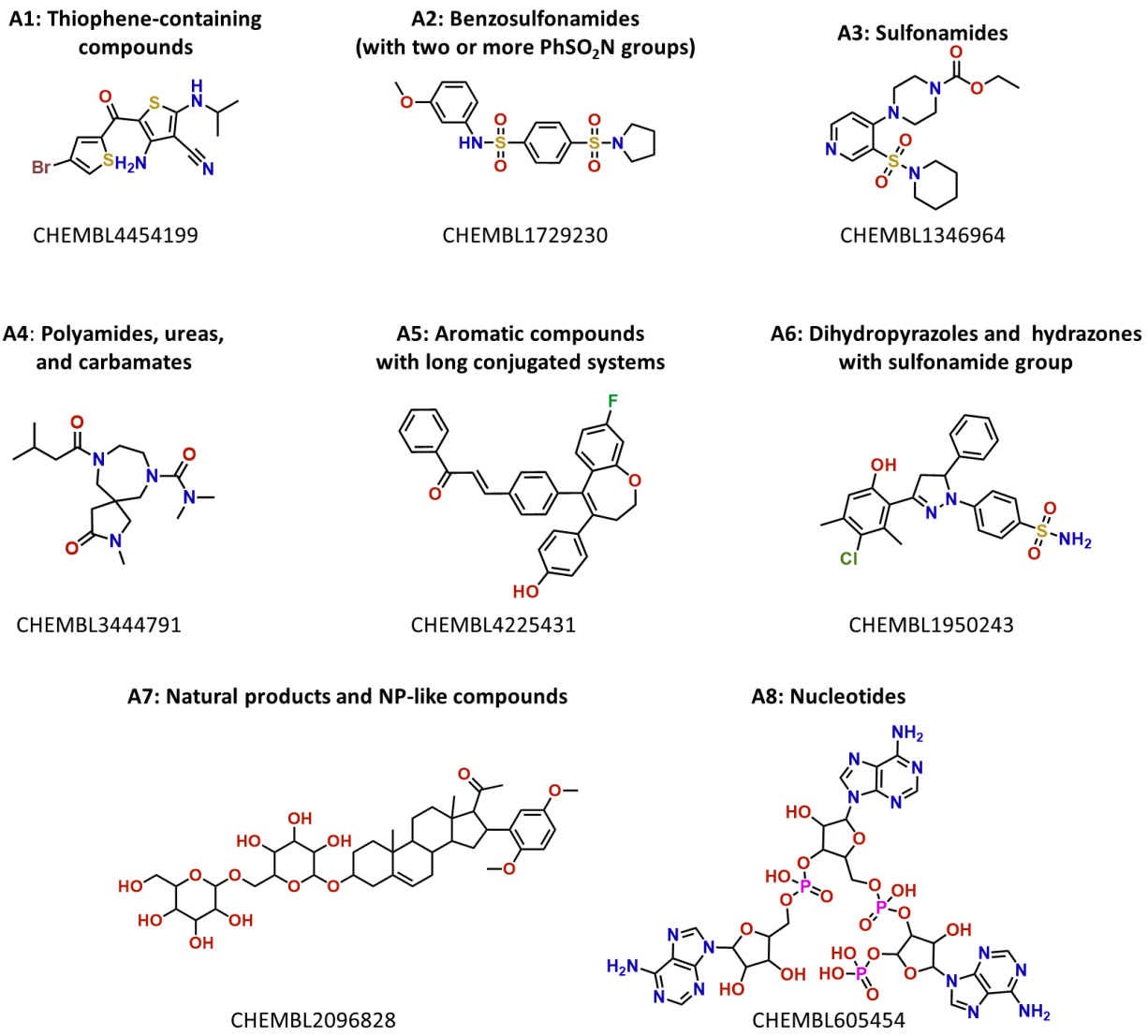

Figure 16. Examples of CHEMBL compounds populating areas from A1 to A8 highlighted in landscapes in Figure 15.

5 Conclusions

In this work, for the first time, the ultra-large chemical space of DNA-encoded libraries (DELs) containing 2,5B compounds in total (2.5K libraries 1M each) was designed and generated using eDesigner and analyzed with the help of GTM. Owing to the probabilistic nature of GTM and efficiency of the libraries analysis and comparison based on the responsibility patterns, it was possible to develop a GTM-based approach for quick selection of DELs occupying the same areas of the chemical space as the reference library. In this work, the goal was to detect the "golden" DEL or "platinum" pool of DELs for primary screening - the libraries containing the highest portion of biologically relevant chemotypes. Therefore, ChEMBL, as the largest database of dose-response activity tests and thus an optimal representation of biologically relevant space, was used as a reference. However, the approach described herein could be applied to any reference library, e.g., actives of a particular biological target.

This approach allowed to identify so-called "platinum" pools of five and three DELs providing the highest coverage of ChEMBL chemical space – 81% and 77%, respectively. Our results suggest that an optimal set for primary screening is the one encompassing three DELs, which, even though containing fewer compounds than in five DELs, still succeeds in covering a large portion of ChEMBL chemical space.

In this project, only a brief structural analysis of DEL chemical space was performed. Without a doubt, a more detailed GTM-based analysis of chemical structures composing DELs and their comparison to ChEMBL and commercially available HTS libraries will improve our understanding of the chemical space accessible via this technology. Further GTM analysis and comparison of generated DELs can be helpful for the enhancement of available BBs libraries and prioritizing some promising synthetic procedures in order to improve the biological relevance of DEL chemical space.

Acknowledgments

The authors are grateful to eMolecules, Inc. for the provided library of commercially available BBs, used for DNA-encoded libraries design.

References

[1] (a) M. S. Attene-Ramos, C. P. Austin, M. Xia, in Encyclopedia of Toxicology, 2014; (b) J. Inglese, D. S. Auld, in Wiley Encyclopedia of Chemical Biology, 2008.
[2] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, G. S. Sittampalam, Nat Rev Drug Discov 2011, 10, 188-195, 10.1038/nrd3368.
[3] R. M. Franzini, C. Randolph, Journal of medicinal chemistry 2016, 59, 6629-6644, 10.1021/acs.jmedchem.5b01874.
[4] N. Favalli, G. Bassi, J. Scheuermann, D. Neri, FEBS letters 2018, 592, 2168-2180,
[5] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, Y. S. Moroz, Iscience 2020, 23, 101681,
[6] S. Brenner, R. A. Lerner, Proc Natl Acad Sci U S A 1992, 89, 5381-5383, 10.1073/pnas.89.12.5381.
[7] R. A. Goodnow Jr, A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery, John Wiley & Sons, 2014.
[8] A. L. Satz, ACS medicinal chemistry letters 2018, 9, 408-410, 10.1021/acsmedchemlett.8b00128.
[9] (a) R. M. Franzini, D. Neri, J. Scheuermann, Accounts of chemical research 2014, 47, 1247-1255, 10.1021/ar400284t; (b) D. Madsen, C. Azevedo, I. Micco, L. K. Petersen, N. J. V. Hansen, Prog Med Chem 2020, 59, 181-249, 10.1016/bs.pmch.2020.03.001.
[10] D. T. Flood, C. Kingston, J. C. Vantourout, P. E. Dawson, P. S. Baran, Israel Journal of Chemistry 2020, 60, 268-280, 10.1002/ijch.201900133.
[11] A. Kontijevskis, J Chem Inf Model 2017, 57, 680-699, 10.1021/acs.jcim.7b00006.
[12] A. Martin, C. A. Nicolaou, M. A. Toledo, Communications Chemistry 2020, 3, 1-9, ARTN 127 10.1038/s42004-020-00374-1.
[13] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Felix, M. P. Magarinos, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Maranon, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, Nucleic Acids Res 2019, 47, D930-D940, 10.1093/nar/gky1075.
[14] C. M. Bishop, M. Svensen, C. K. I. Williams, Neural Computation 1998, 10, 215-234, Doi 10.1162/089976698300017953.
[15] Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, A. Varnek, J Chem Inf Model 2021, 61, 179-188, 10.1021/acs.jcim.0c00936.
[16] I. Casciuc, Y. Zabolotna, D. Horvath, G. Marcou, J. r. Bajorath, A. Varnek, Journal of chemical information and modeling 2018, 59, 564-572, 10.1021/acs.jcim.8b00650.
[17] F. W. Goldberg, J. G. Kettle, T. Kogej, M. W. D. Perry, N. P. Tomkinson, Drug Discovery Today 2015, 20, 11-17, 10.1016/j.drudis.2014.09.023.
[18] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, Mol Inform 2010, 29, 855-868, 10.1002/minf.201000099.
[19] K. Klimenko, G. Marcou, D. Horvath, A. Varnek, J Chem Inf Model 2016, 56, 1438-1454, 10.1021/acs.jcim.6b00192.
[20] Y. Zabolotna, D. M. Volochnyuk, S. V. Ryabukhin, K. Gavrylenko, D. Horvath, O. Klimchuk, O. Oksiuta, G. Marcou, A. Varnek, ChemRxiv. Cambridge: Cambridge Open Engage 2021, 10.33774/chemrxiv-2021-v53hl-v2
[21] LillyMol: Eli Lilly Computational Chemistry and Chemoinformatics Group Toolkit, 2020, https://github.com/EliLillyCo/LillyMol
[22] D. Horvath, G. Marcou, A. Varnek, Drug Discov Today Technol 2019, 32-33, 99-107, 10.1016/j.ddtec.2020.06.003.
[23] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, J Comput Aided Mol Des 2015, 29, 1087-1108, 10.1007/s10822-015-9882-z.
[24] eMolecules Inc., https://www.emolecules.com/
[25] Enamine Ltd., https://enamine.net/
[26] ChemaAxon. JChem, 2020, Version 20.8.3, ChemAxon Ltd: Budapest, Hungary,
[27] Virtual Screening Web Server 2020, http://infochim.u-strasbg.fr/webserv/VSEngine.html
[28] C. Zambaldo, S. N. Geigle, A. L. Satz, Org Lett 2019, 21, 9353-9357, 10.1021/acs.orglett.9b03553.
[29] A. L. Satz, J. Cai, Y. Chen, R. Goodnow, F. Gruber, A. Kowalczyk, A. Petersen, G. Naderi-Oboodi, L. Orzechowski, Q. Strebel, Bioconjug Chem 2015, 26, 1623-1632, 10.1021/acs.bioconjchem.5b00239.
[30] G. Landrum, RDKit: Open-Source Cheminformatics Software, 2016, http://www.rdkit.org/, https://github.com/rdkit/rdkit
[31] C. A. Lipinski, Journal of pharmacological and toxicological methods 2000, 44, 235-249, 10.1016/S1056-8719(00)00107-6.
[32] M. P. Gleeson, Journal of medicinal chemistry 2008, 51, 817-834, 10.1021/jm701122q.
[33] (a) O. Eidam, A. L. Satz, MedChemComm 2016, 7, 1323-1331, 10.1039/C6MD00221H; (b) Z. Wu, T. L. Graybill, X. Zeng, M. Platchek, J. Zhang, V. Q. Bodmer, D. D. Wisnoski, J. Deng, F. T. Coppo, G. Yao, ACS combinatorial science 2015, 17, 722-731, 10.1021/acscombsci.5b00124.

Received: ((will be filled in by the editorial staff))
Accepted: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff