Exploration of the chemical space of DNA-encoded libraries

Yuliana Zabolotna[1], Regina Pikalyova[1], Dmitriy M.Volochnyuk[3,4], Dragos Horvath[1], Gilles Marcou[1], Alexandre Varnek[1,2]*

Abstract: DNA-Encoded Library (DEL) technology has emerged as an alternative method for bioactive molecule discovery in medicinal chemistry. It enables 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 theultra-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

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, 2]. Since its emergence in the 1990s, HTS has delivered numerous lead molecules for drug development[3]. Nevertheless, this technology has several limitations, such as expensive robotic equipment and compound libraries, that are available mostly to large pharmaceutical companies[4]. The number of compounds that can be screened in one HTS campaign is usually limited to a million[5], while the chemical space of synthetically accessible molecules is far larger[6].

DNA-encoded library (DEL) technology has partially solved these problems[7]. It consists of the creation of ultra-large libraries of DNA-encoded compounds using water-based combinatorial chemistry and their screening against soluble target proteins using binding affinity selection[8]. 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-

1. *University of Strasbourg, Laboratoire de Chemoinformatique , 4, rue B. Pascal, Strasbourg 67081 (France) *e-mail: 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. *Enamine Ltd. 78 Chervonotkatska str., 02660 Kiev, Ukraine*

and-pool approach[9] and thus allow to produce chemically versatile libraries of enormous size[10, 11]. DEL compounds are screened all at once in a single vessel in contrast to individual compound screening in HTS[8]. Simple experimental setup of affinity selection accessible both in industry and university laboratories allows cheap and fast hits identification.[12] Many successful stories of employing this technology were published, including DEL-derived hits that progressed to clinic[9].

However, up to this point, most efforts were focused on the analysis of the libraries of BBs or identified active compounds[4]. 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[13]. 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[14]. 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 generated DEL space to the chemical space of biologically relevant ChEMBL[15] compounds using Generative Topographic Mapping (GTM) – a very efficient dimensionality reduction method[16]. GTM has proved to be a powerful tool for "Big Data" analysis and visualization (up to 1B compounds)[17]. 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.[18] Only one of the several uMaps (uMap1, see corresponding article) has been used in this study for simplicity, but the study could be extended to consensus mapping on several uMaps.

METHODS

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)[19] 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[20] 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)[21]. To achieve even better coverage, complementary DELs were added to the "golden" one to give a "platinum" pool of DELs.

BBs selection

Before DEL design and generation, input BBs were filtered according to Ro2 with the help of SynthI[22]. Ro2 is a guideline to choose high-quality BBs that can give access to drug-like molecules[19]. 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.

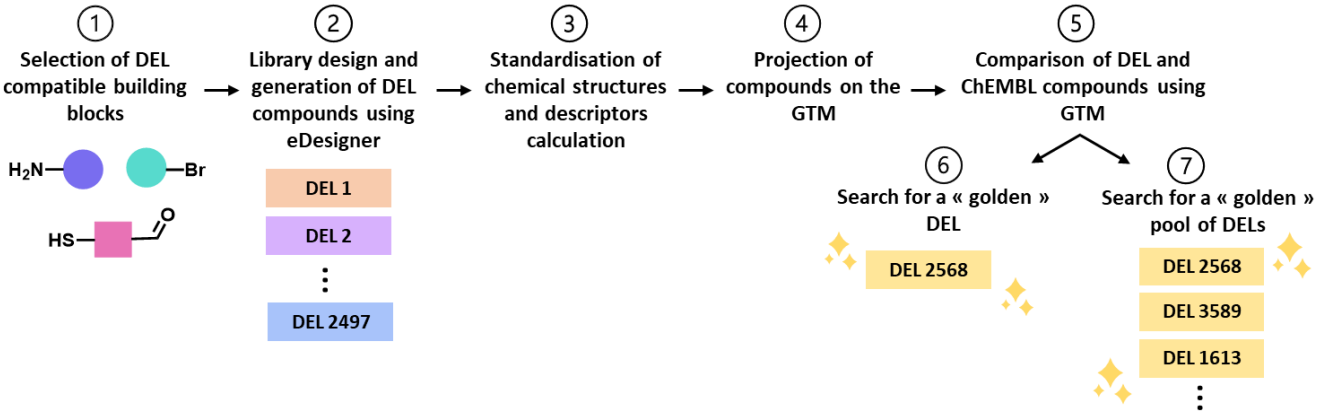


Figure 1. Workflow of the project. The rectangles represent separa te DNA encoded libraries (DELs).

DEL generation with eDesigner

For the generation of chemical space of DELs, the eDESIGNER[14] 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[14]) 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.[23] 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.

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[16] 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

numbers 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.

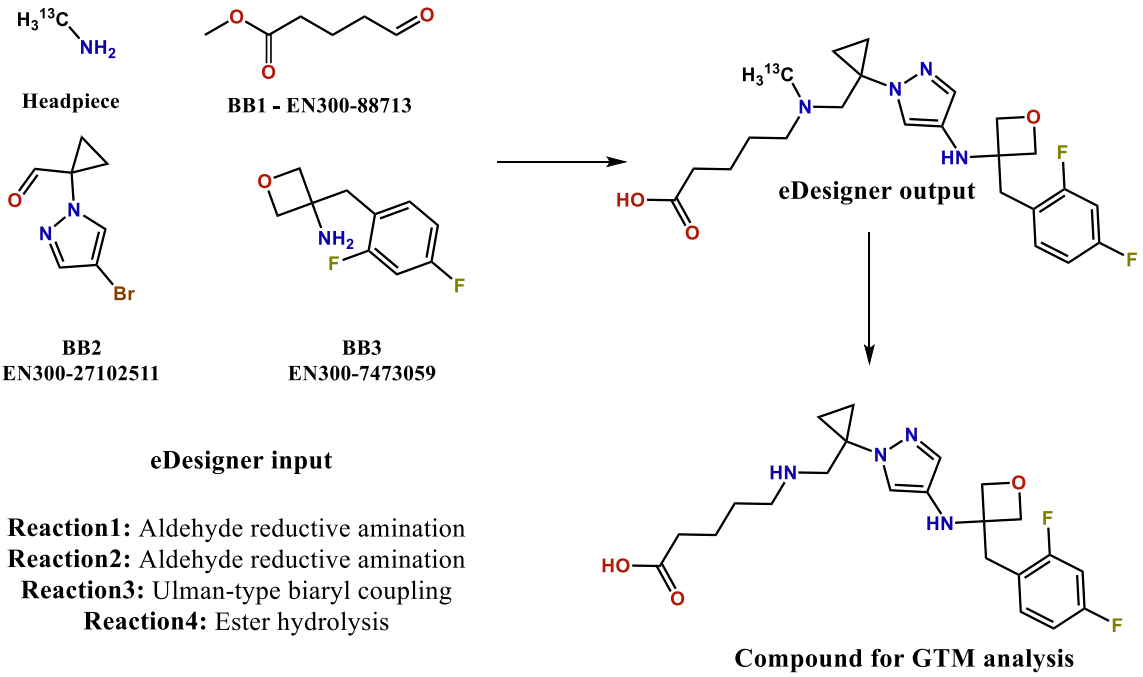


Figure 2. Example of DEL compound generation by eDesigner. The user should provide headpiece and the list of BBs; an appropriate list of reactions will be selected automatically by eDesigner, and respective compounds are generated. The isotopic mark is placed by eDesigner in order to know the position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

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 (library1) and ChEMBL (library2) 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[24].

Universal GTM

The concept of Universal GTM (UGTM) was introduced by Sidorov et al.[25] and further developed by Casciuc et al.[18] 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[20]. 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[15], 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[17].

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.[26] 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[24]. 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 with ChEMBL, i.e. to evaluate the proportion of ChEMBL RPs ("structural motifs") also covered by a given DEL.

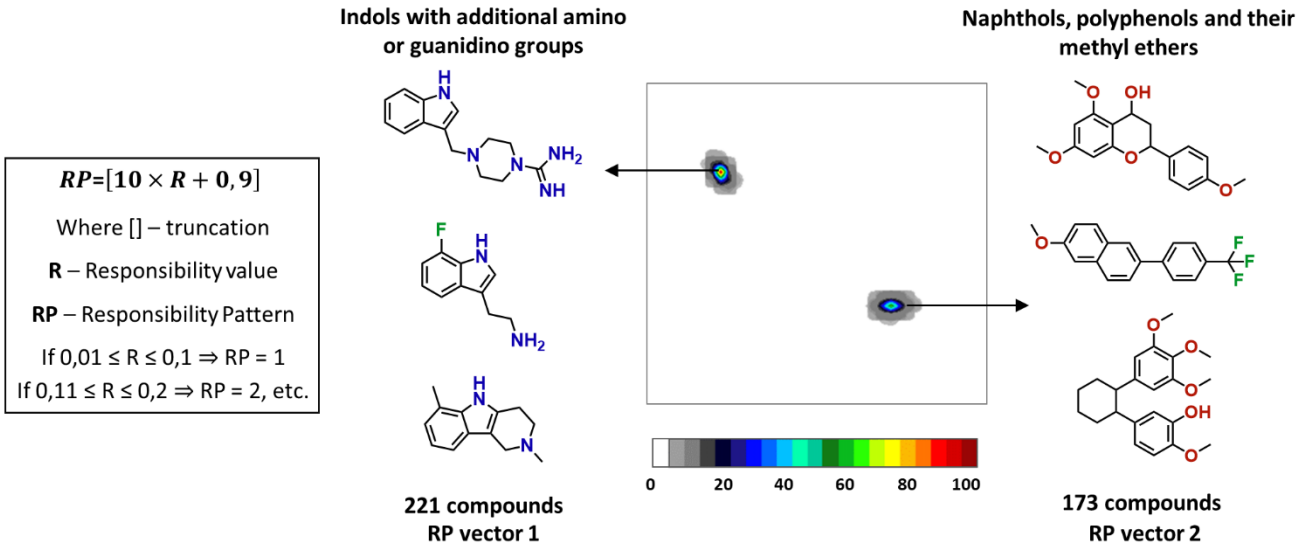


Figure 3. Left: formula for responsibility pattern (RP) calculation. Right: example of compounds sharing the same RPs and their position on the density landscape - a map colored by local density of compounds. Highly populated zones are colored in red, underpopulated ones - in grey.

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:

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

DATA

Commercially available BBs

A set of 450K commercially available BBs was provided by eMolecules Inc[27]. They were complemented by an "orthogonal" (i.e., containing completely different BBs) dataset of 10K Enamine[28] in-stock BBs. Among them, only 79,141 BBs that satisfy Ro2 and eDesigner build-in DNA-compatibility filters were selected.

ChEMBL (biologically tested compounds)

ChEMBL is a database containing >2M diverse and biologically relevant compounds against >14K biological targets[15]. 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[29] according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics in the University of Strasbourg.[30]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.

RESULTS AND DISCUSSION

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[31]. For example, only a few heterocyclisations optimized for DEL synthesis were described, such as benzimidazole,

imidazolidinone, thiazole synthesis, and some others[32]. Nevertheless, even a few reactions can give rise to structurally diverse DELs if abundant building blocks (BBs) sets are employed for their generation.

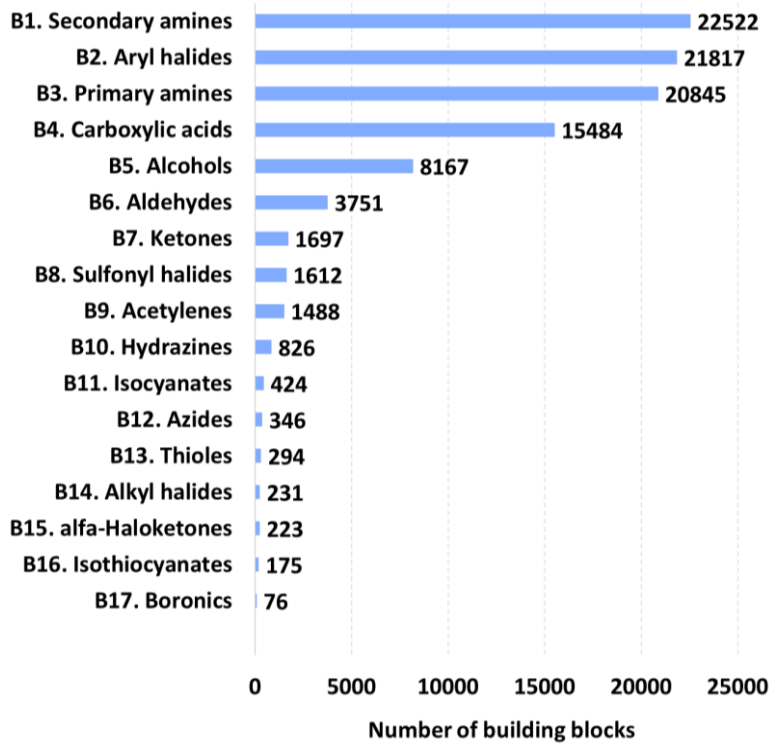


Figure 4. Monofunctional DNA-compatible commercially available BBs.

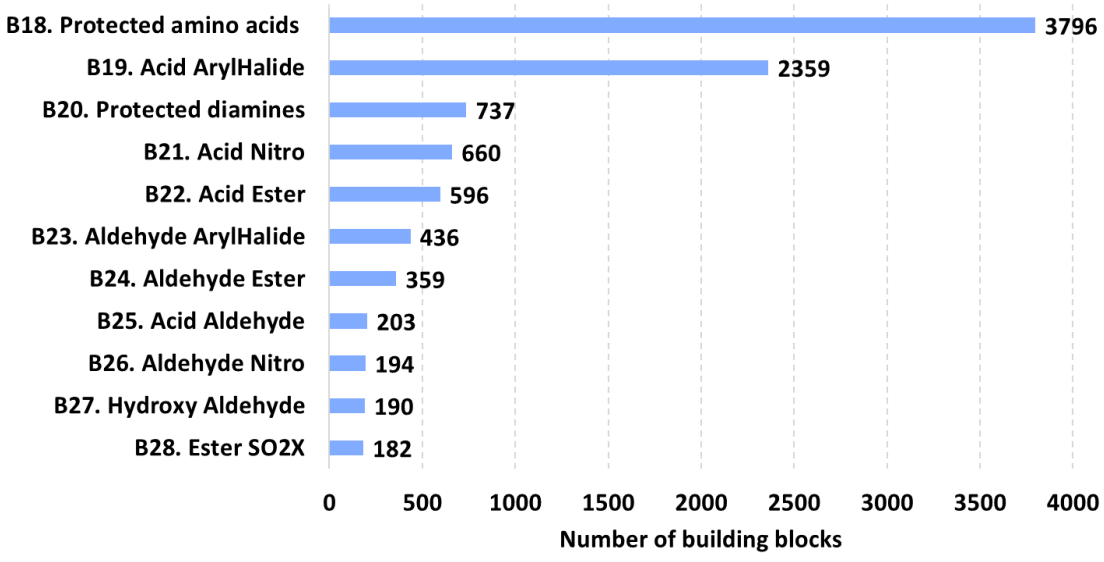


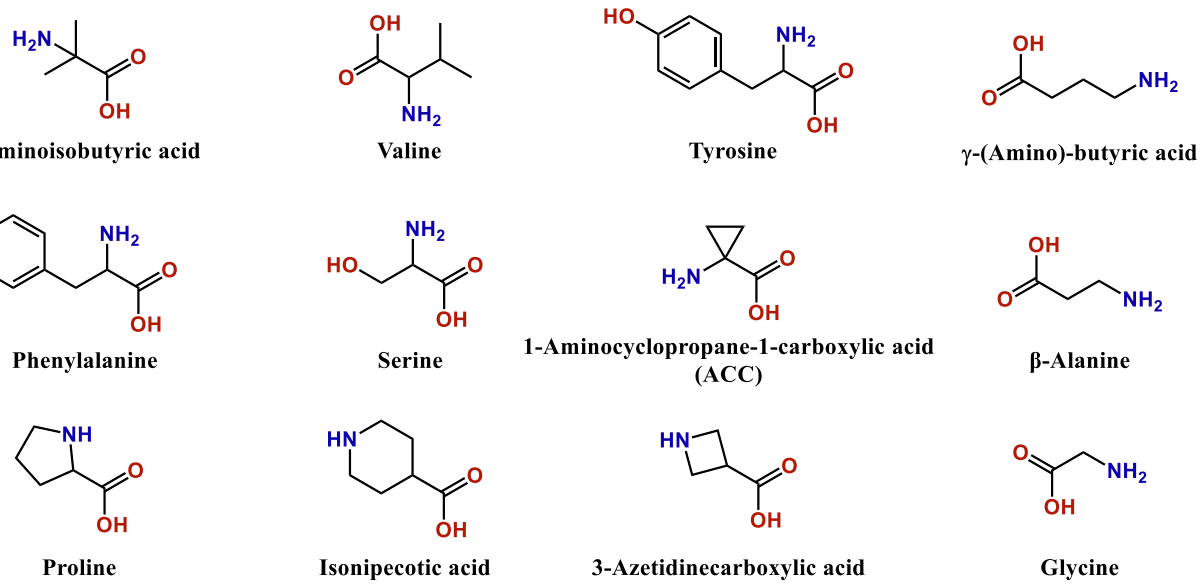
Figure 5. Bifunctional DNA-compatible commercially available BBs.

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**). 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 being 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 only 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 combinations in the BB library.

Amino acids with the highest number of protecting group variations in the commercially available libraries



Diamines with the highest number of protecting group variations in the commercially available libraries

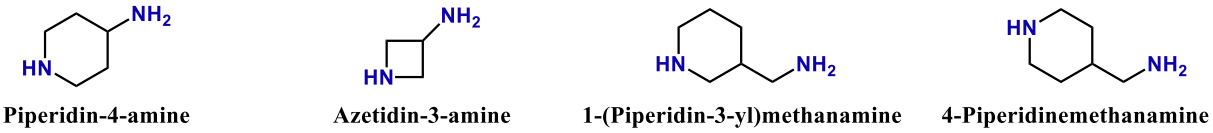


Figure 6. AA (I) and diamines (II), represented in the commercially available libraries of DNA-compatible BBs with the highest number of protected variations (N-Boc, N-Fmoc, various esters etc.)

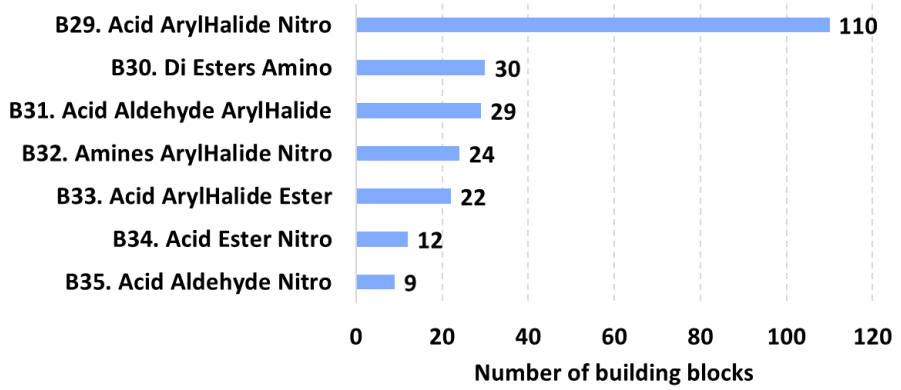


Figure 7. Trifunctional DNA-compatible commercially available BBs.

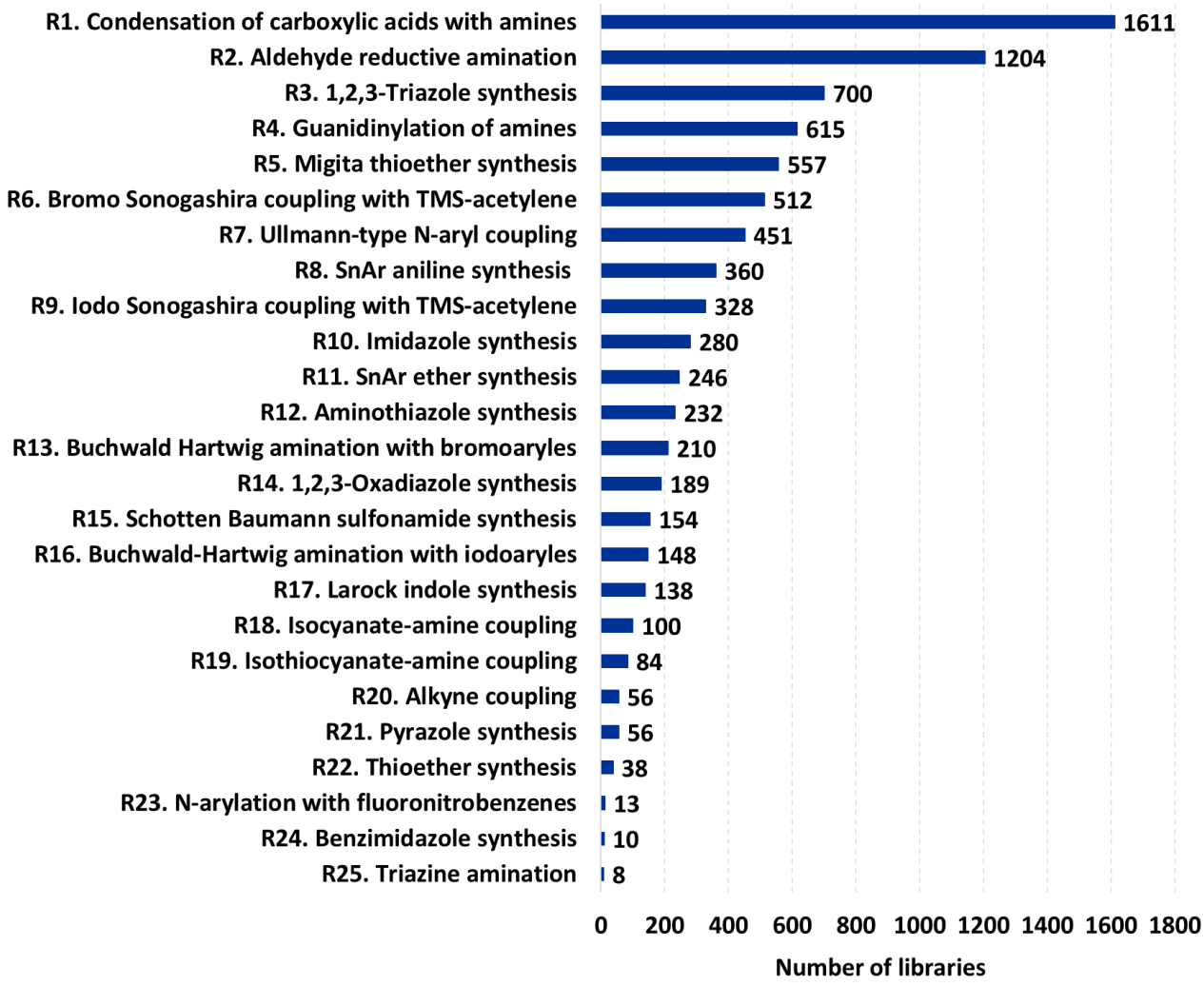


Figure 8. Frequency of the use of a particular reaction in DELs generation.

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 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 observed 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, 2,495 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**. 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 has been enumerated for one of the DELs, and density landscapes have been 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 large DELs for GTM-based analysis.

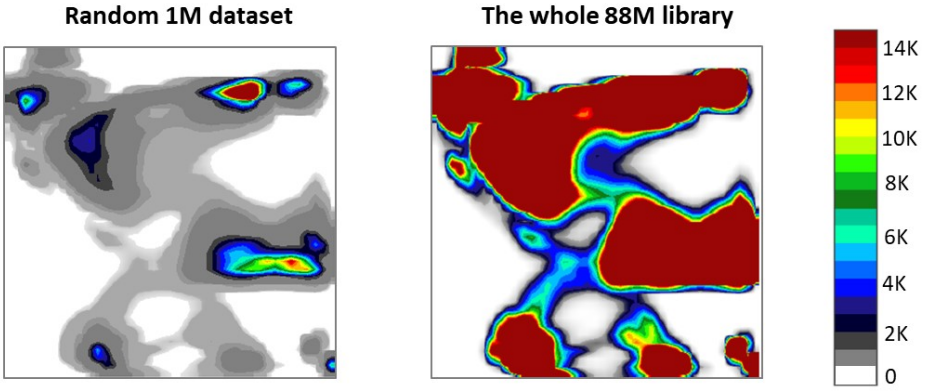


Figure 9. Comparison of the density distribution for the 1M randomly generated compounds and the whole DEL(88M). The color scale encodes the corresponding number of compounds residing in each colored node of the map.

Physicochemical properties of generated libraries

Out of total 2,495 generated DELs, 77 are produced by a single coupling reaction of 2 BBs (hence the label "2BB libraries"). The remaining 2,418 DELs are "3BB libraries". The physicochemical properties were calculated using RDKit[33]. Drug-like[34] (MW ≤ 500; LogP ≤ 5; the number of H-bond donors ≤ 5; the number of H-bond acceptors ≤ 10; ring counts ≤ 10) and lead-like[35] (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.

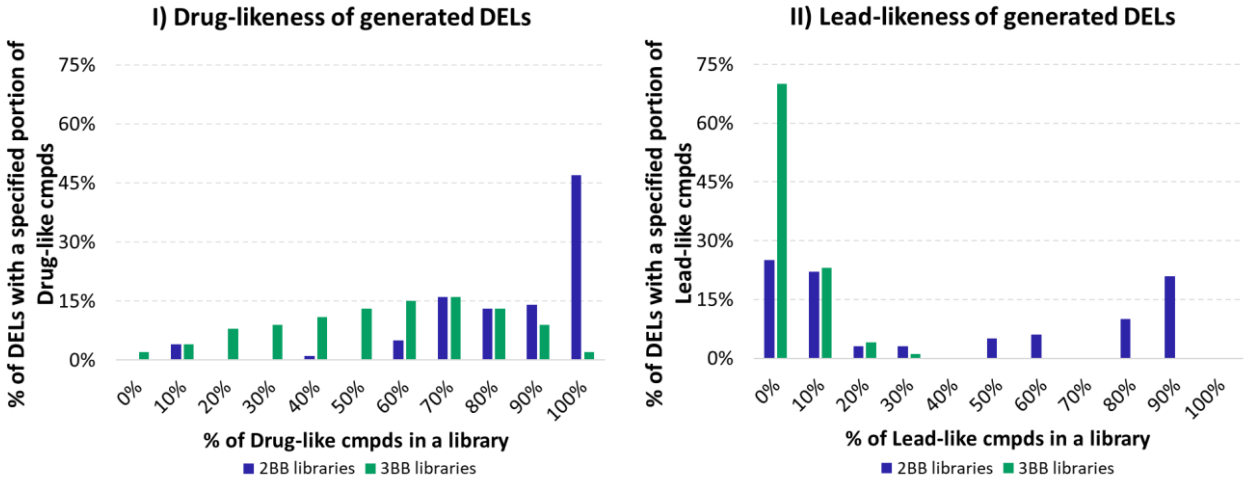


Figure 10. Comparison of (I) drug- and (II) lead-likeness of 2BB and 3BB libraries: percentage of 2BB and 3BB libraries having a particular portion of compounds satisfying respective filters is given.

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.

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 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 DEL2568 with 60% 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 multicolored 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. However, DEL2568 is the leader among all 2,5K DELs - multicolored 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.

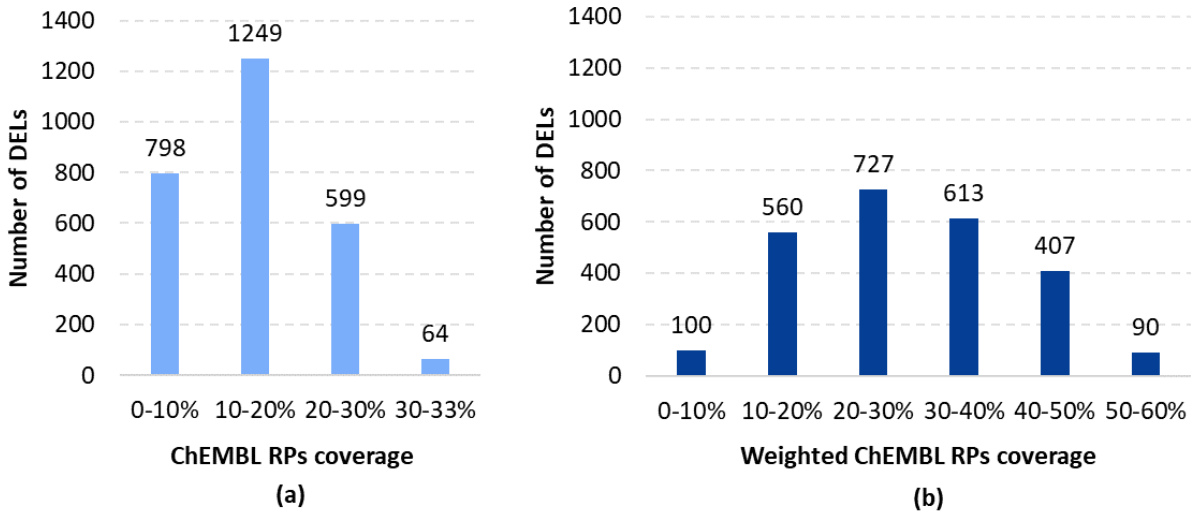


Figure 11. (a) Number of DELs with different coverage of ChEMBL responsibility patterns (RPs) (b) Number of DELs with different percentages of ChEMBL RPs coverage weighted by the RPs population (number of ChEMBL compounds per RP).

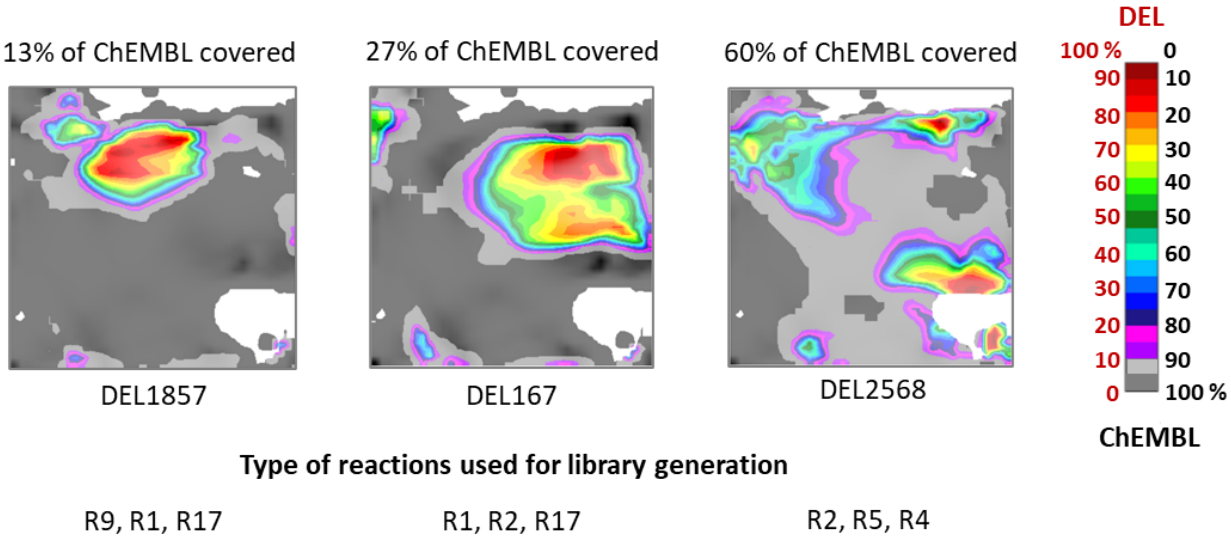


Figure 12. Class landscapes comparing a particular DEL with ChEMBL. From left to right: comparison of ChEMBL to DEL1857, DEL167, and DEL2568. Dark grey zones are populated exclusively by ChEMBL compounds, while all other colors indicate areas also containing DEL compounds in a different ratio. White regions correspond to the empty areas of the chemical space. Below each landscape, , a library ID and IDs for corresponding reaction types are given.

There are around 60 libraries with similar chemical space coverage and diversity, but here, we will limit the discussion to the DEL2568 as an example of a "golden" DEL. 88 Million compounds from this DEL can be obtained by sequentially employing three reactions: aldehyde reductive amination, Migita thioether synthesis, and guanidinylation of amines (see **Figure 14**, DEL2568). BBs used for this DEL design are three aromatic mercaptoaldehydes, 8,914 aryl bromides, and 3,311 amines. As was discussed earlier, the last two are the classes with the highest number of diverse BBs (**Figure 4**). Therefore, a random selection of BBs for DEL generation from such various and numerous collections results in higher coverage of ChEMBL chemical space. DEL2568 was chosen here as an example of a "golden" library because it outruns all

other libraries by 3% of weighted ChEMBL coverage, corresponding to approximately 45K of biologically relevant compounds. However, if the presence of thioether or guanidine groups is not desirable, there is still a diverse choice of DELs that do not contain such moieties.

Search for the "platinum" set of DELs

As shown on the class landscape for DEL2568 in **Figure 11**, 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[36, 37] was considered. According to it, several distinct DELs may be further combined to create another more complex mixture, called "library pool", which can then be simultaneously screened. 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.

To achieve that, first of all, 64 DELs that have the highest coverage of ChEMBL RPs were chosen. Each of these DELs 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 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. 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 worth it. Therefore, it is irrational to use a pool of DELs composed of more than five libraries.

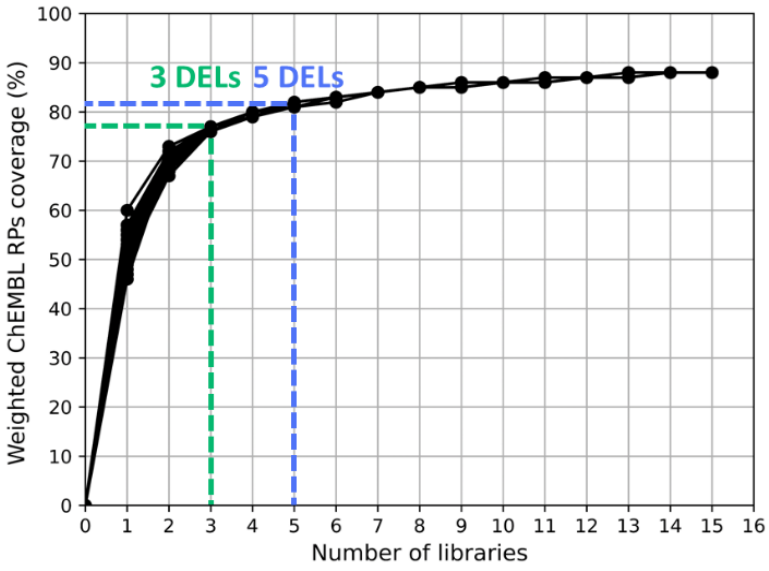


Figure 13. The percentage of the ChEMBL coverage, weighted by the number of compounds sharing common RPs, as a function of the number of libraries in the set. Green and blue dashed lines highlight the points for three and five DELs.

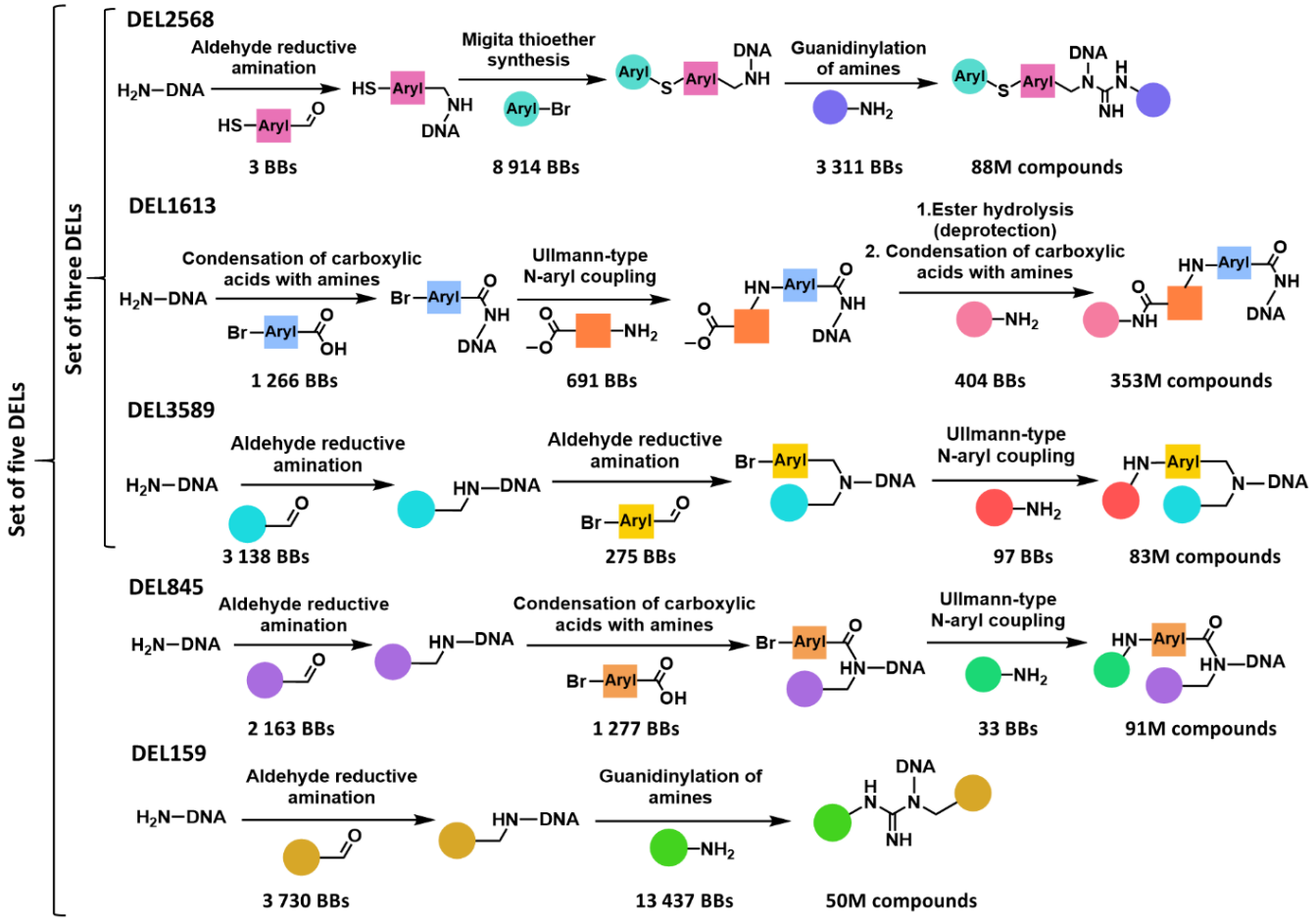


Figure 14. Reactions and BBs required for synthesis of the "golden" DEL and libraries composing "platinum" pools of libraries.

If described above DEL2568 is used as a starting DEL, the "platinum" pool of five DELs will be composed of such libraries: DEL2568, DEL1613, DEL159, DEL845, and DEL3589. Overall, they contain 665M compounds. Reactions used for the generation of these five DELs are shown in **Figure 14:** aldehyde reductive amination (R2), Migita thioether synthesis (R5), Ullmann-type N-aryl coupling (R7), condensation of carboxylic acids with amines (R1), and guanidinylation of amines (R4). 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 (DEL2568, DEL1613, DEL3589) can be even more convenient since it contains fewer compounds (524M) and yet still allows to cover a large portion of ChEMBL (78%).

The physicochemical properties of the selected libraries have been calculated and analyzed (**Table 1**). It appears that half of DEL2568 compounds are drug-like, while the portion of lead-like molecules is almost negligible. Complementary DELs forming a "platinum" pools of three and five DELs possess higher drug- and lead-likeness, which influenced the number of corresponding compounds. Indeed, the percentage of drug-like compounds is increasing for the pool of 3 DELs (60.8%) and even more so in the case of 5 DELs (70.4%). Likewise, the portion of lead-like compounds peaks at 21% for the pool of 5 DELs.

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. On the landscape of the "platinum" pool of three DELs, the ChEMBL areas from A1 to A7 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)**).

Table 1. The portion of drug-like and lead-like compounds in the selected "golden" DEL and "platinum" pools of three and five DELs.

	Portion of drug-like compounds	Portion of lead-like compounds
"Golden" DEL2568	50%	1.5%
"Platinum" pool of 3 DELs	60.8%	6.2%
"Platinum" pool of 5 DELs	70.4%	21.7%

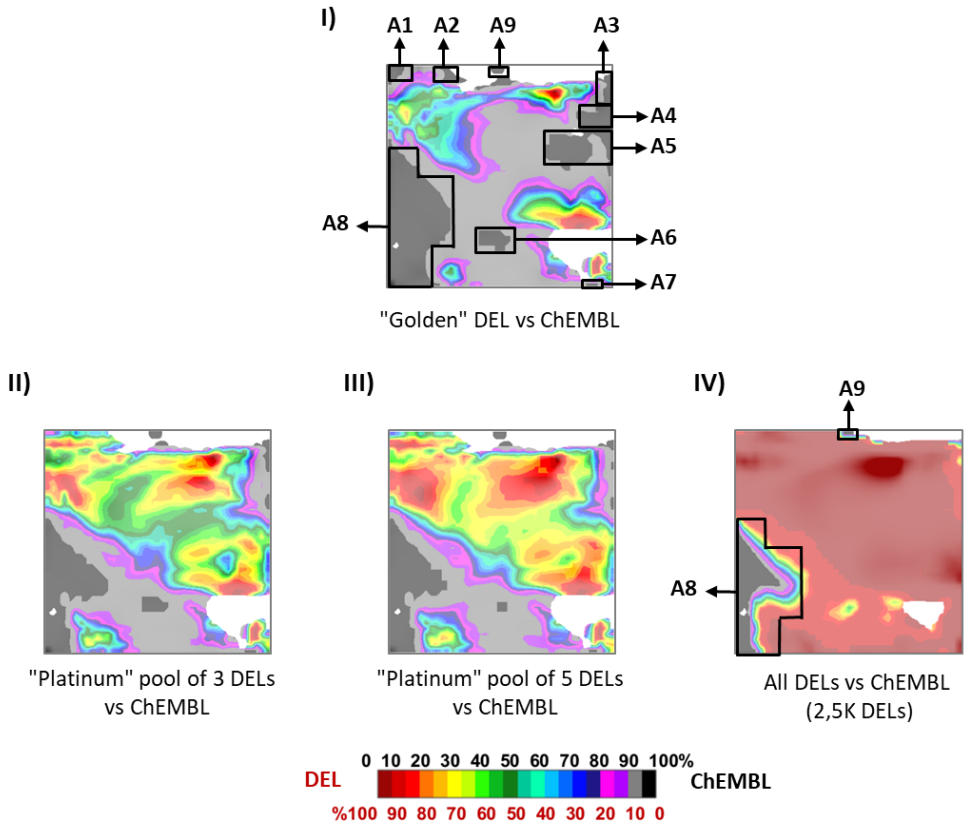
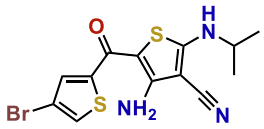


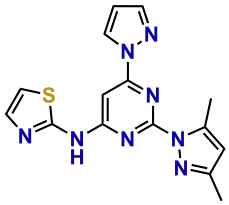
Figure 15. Comparison of ChEMBL and I) "golden" DEL, II) a pool of three DELs, III) a pool of five DELs, and IV) all 2,5K DELs. Multicolored zones are populated by both ChEMBL and DEL compounds, dark grey zones – only by ChEMBL compounds. White regions correspond to the empty areas of the chemical space. Examples of compounds populating highlighted areas A1-A9 are provided in **Figure 16**

A1: Thiophene-containing compounds



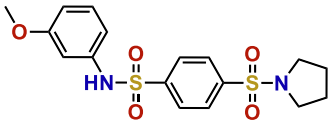
CHEMBL4454199

A2: Thiazoles and thiadiazoles



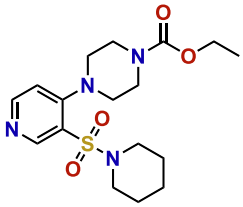
CHEMBL3100167

A3: Benzosulfonamides (with two or more PhSO2N groups)



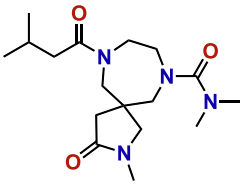
CHEMBL1729230

A4: Sulfonamides



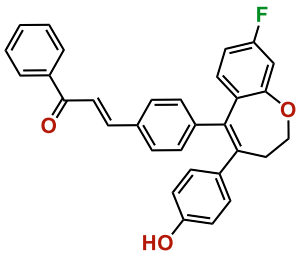
CHEMBL1346964

A5: Polyamides, ureas, and carbamates



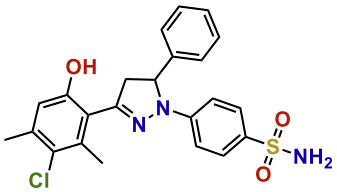
CHEMBL3444791

A6: Aromatic compounds with long conjugated systems



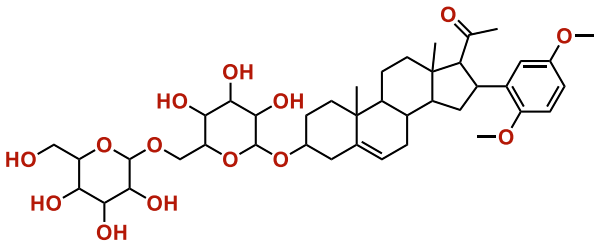
CHEMBL4225431

A7: Dihydropyrazoles and hydrazones with sulfonamide group



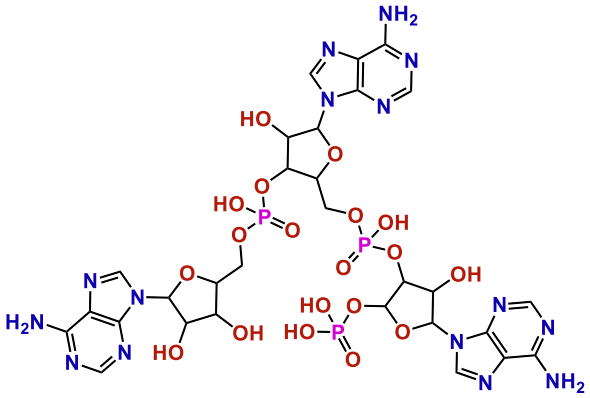
CHEMBL1950243

A8: Natural products and NP-like compounds



CHEMBL2096828

A9: Nucleotides



CHEMBL605454

Figure 16. Examples of CHEMBL compounds populating areas from A1 to A9 highlighted in landscapes in **Figure 15**.

However, neither three nor five libraries succeeded in covering areas A8 and A9 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 A8 and A9 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.

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 a 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 the so-called "platinum" pools of five and three DELs providing the highest coverage of ChEMBL chemical space – 82% and 78%, 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. Analysis of physicochemical properties of the "golden" DEL revealed that half of the compounds are drug-like, and in the case of the pool of 3 DELs, this percentage rises to 60%. The portion of lead-like molecules, however, is negligible.

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.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Attene-Ramos, M. S.; Austin, C. P.; Xia, M. High Throughput Screening. In *Encyclopedia of Toxicology*, Wexler, P., Ed.; Academic Press: Oxford, 2014, pp 916-917.
2. Inglese, J.; Auld, D. S., High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. *Wiley Encyclopedia of Chemical Biology* **2008**, 1-15.
3. Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S., Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011**, 10, 188-95.

4. Franzini, R. M.; Randolph, C., Chemical Space of DNA-Encoded Libraries. *J. Med. Chem.* **2016**, 59, 6629-44.
5. Favalli, N.; Bassi, G.; Scheuermann, J.; Neri, D., DNA-encoded chemical libraries–achievements and remaining challenges. *FEBS Lett.* **2018**, 592, 2168-2180.
6. 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.
7. Brenner, S.; Lerner, R. A., Encoded combinatorial chemistry. *Proc Natl Acad Sci U S A* **1992**, 89, 5381-3.
8. Goodnow Jr, R. A., *A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery*. John Wiley & Sons: 2014.
9. Satz, A. L., What do you get from DNA-encoded libraries? *ACS medicinal chemistry letters* **2018**, 9, 408-410.
10. Franzini, R. M.; Neri, D.; Scheuermann, J., DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. *Acc. Chem. Res.* **2014**, 47, 1247-55.
11. Madsen, D.; Azevedo, C.; Micco, I.; Petersen, L. K.; Hansen, N. J. V., An overview of DNA-encoded libraries: A versatile tool for drug discovery. *Progress in medicinal chemistry* **2020**, 59, 181-249.
12. Flood, D. T.; Kingston, C.; Vantourout, J. C.; Dawson, P. E.; Baran, P. S., DNA Encoded Libraries: A Visitor's Guide. *Isr. J. Chem.* **2020**, 60, 268-280.
13. Kontijevskis, A., Mapping of drug-like chemical universe with reduced complexity molecular frameworks. *J. Chem. Inf. Model.* **2017**, 57, 680-699.
14. Martín, A.; Nicolaou, C. A.; Toledo, M. A., Navigating the DNA encoded libraries chemical space. *Commun. Chem.* **2020**, 3, 127.
15. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M., ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, 47, D930-D940.
16. Bishop, C. M.; Svensen, M.; Williams, C. K. I., GTM: The generative topographic mapping. *Neural Comput.* **1998**, 10, 215-234.
17. Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A., Chemography: Searching for Hidden Treasures. *J Chem Inf Model* **2021**, 61, 179-188.
18. Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A., Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J Chem Inf Model* **2019**, 59, 564-572.
19. Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W.; Tomkinson, N. P., Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discov. Today* **2015**, 20, 11-7.
20. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, 29, 855-68.
21. Sidorov, P.; Viira, B.; Davioud-Charvet, E.; Maran, U.; Marcou, G.; Horvath, D.; Varnek, A., QSAR modeling and chemical space analysis of antimalarial compounds. *J. Comput. Aided Mol. Des.* **2017**, 31, 441-451.
22. Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Gavrylenko, K.; Horvath, D.; Klimchuk, O.; Oksiuta, O.; Marcou, G.; Varnek, A., SynthI: a new open-source tool for synthon-based library design *ChemRxiv. Cambridge: Cambridge Open Engage;* **2021**, doi: 10.33774/chemrxiv-2021-v53hl-v2. This content is a preprint and has not been peer-reviewed.
23. *LillyMol: Eli Lilly Computational Chemistry and Chemoinformatics Group Toolkit.* https://github.com/EliLillyCo/LillyMol 2020.
24. Horvath, D.; Marcou, G.; Varnek, A., Generative topographic mapping in drug design. *Drug Discov Today Technol* **2019**, 32-33, 99-107.
25. Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D., Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided Mol. Des.* **2015**, 29, 1087-108.
26. Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A., Chemical Space Mapping and Structure-Activity Analysis of the ChEMBL

Antiviral Compound Set. *J Chem Inf Model* **2016**, 56, 1438-54.
27. *eMolecules, Inc.* https://www.emolecules.com/.
28. *Enamine, Ltd.* https://enamine.net/.
29. *ChemAxon. JChem, Version 20.8.3, ChemAxon, Ltd: Budapest, Hungary* **2020**.
30. *Virtual Screening Web Server.* http://infochim.u-strasbg.fr/webserv/VSEngine.html, December 2020.
31. Zambaldo, C.; Geigle, S. N.; Satz, A. L., High-Throughput Solid-Phase Building Block Synthesis for DNA-Encoded Libraries. *Org. Lett.* **2019**, 21, 9353-9357.
32. Satz, A. L.; Cai, J.; Chen, Y.; Goodnow, R.; Gruber, F.; Kowalczyk, A.; Petersen, A.; Naderi-Oboodi, G.; Orzechowski, L.; Strebel, Q., DNA Compatible Multistep Synthesis and Applications to DNA Encoded Libraries. *Bioconjug Chem* **2015**, 26, 1623-32.
33. Landrum, G., *RDKit: Open-Source Cheminformatics Software.* http://www.rdkit.org.
34. Lipinski, C. A., Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, 44, 235-49.
35. Gleeson, M. P., Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, 51, 817-34.
36. Eidam, O.; Satz, A. L., Analysis of the productivity of DNA encoded libraries. *MedChemComm* **2016**, 7, 1323-1331.
37. Wu, Z.; Graybill, T. L.; Zeng, X.; Platchek, M.; Zhang, J.; Bodmer, V. Q.; Wisnoski, D. D.; Deng, J.; Coppo, F. T.; Yao, G.; Tamburino, A.; Scavello, G.; Franklin, G. J.; Mataruse, S.; Bedard, K. L.; Ding, Y.; Chai, J.; Summerfield, J.; Centrella, P. A.; Messer, J. A.; Pope, A. J.; Israel, D. I., Cell-Based Selection Expands the Utility of DNA-Encoded Small-Molecule Library Technology to Cell Surface Drug Targets: Identification of Novel Antagonists of the NK3 Tachykinin Receptor. *ACS Combinatorial Science* **2015**, 17, 722-731.