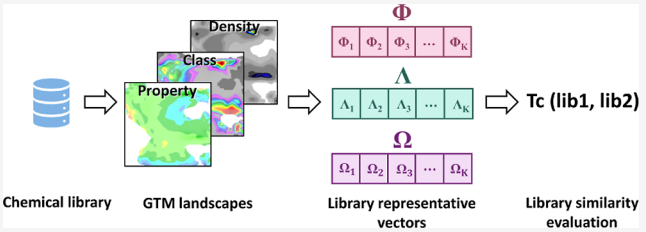
Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case

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

Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case

Regina Pikalyova, Yuliana Zabolotna, Dragos Horvath, Gilles Marcou, and Alexandre Varnek*

ABSTRACT: The development of DNA-encoded library (DEL) technology introduced new challenges for the analysis of chemical libraries. It is often useful to consider a chemical library as a standalone chemoinformatic object—represented both as a collection of independent molecules, and yet an individual entity—in particular, when they are inseparable mixtures, like DELs. Herein, we introduce the concept of chemical library space (CLS), in which resident items are individual chemical libraries. We define and compare four vectorial library representations obtained using generative topographic mapping. These allow for an effective comparison of libraries, with the ability to tune and chemically interpret the similarity relationships. In particular, property-tuned CLS encodings enable us to simultaneously compare libraries with respect to both property and chemotype distributions. We apply the various CLS encodings for the selection problem of DELs that optimally "match" a reference collection (here ChEMBL28), showing how the choice of the CLS descriptors may help to fine-tune the "matching" (overlap) criteria. Hence, the proposed CLS may represent a new efficient way for polyvalent analysis of thousands of chemical libraries. Selection of an easily accessible compound collection for drug discovery, as a substitute for a difficult to produce reference library, can be tuned for either primary or target-focused screening, also considering property distributions of compounds. Alternatively, selection of libraries covering novel regions of the chemical space with respect to a reference compound subspace may serve for library portfolio enrichment.



1. INTRODUCTION

Chemical library design and evaluation have always been one of the central aspects of computer-aided drug design. Over the last decades, the main efforts in chemoinformatics were directed toward different ways of chemical structure encoding, various approaches for chemical space representation, visualization, and efficient ways to characterize the chemical composition of analyzed collections. Considering that at the time medicinal chemists were operating with only a few compound collections, a given library (in-house stock or preferable supplier catalog) was a space of exploration, and underlying compounds were the objects in this analysis. Later on, advances in organic chemistry (*e.g.*, parallel synthesis) increased significantly the number of distinct chemical collections, and the compound population in those libraries exploded, especially so for tangible libraries. However, the association of a given molecule to a "classical" compound library was still somewhat arbitrary—one collection could be enhanced using compounds from the other or even a new library could be created by cherry-picking compounds from numerous different collections. Moreover, considering that each compound was synthesized and biologically tested separately, it was logical to only evaluate libraries at the level of individual molecules.

With time, combinatorial chemistry has advanced to the point that it is now possible to simultaneously synthesize a mixture containing millions of compounds in a few simple and easily automatable steps. A variety of encoding methods have been developed, enabling the recording of specific reaction rules and building block (BB) combinations defining a mixture.[1] Affinity selection combined with decoding techniques allowed for the simultaneous biological screening of ultra-large compound collections contained within a single Eppendorf tube. It is from the background of these advancements that DNA-encoded library (DEL) technology emerged and recently became an attractive tool for hit identification successfully applied at the early stages of drug discovery.[2,3] DEL technology enables much faster and cheaper identification of potential hits as opposed to widely used but quite expensive high-throughput screening. DEL technology is associated with various challenges—both experimental and computational. One of them is related to the fact that a library of DNA-encoded molecules is synthesized and tested as a whole. It can, of course, be designed by thorough choice of its

BBs or pooling multiple DELs together—but, once the mixture is produced, it cannot be broken down to individual molecules any longer. This means, it is impossible to exclude or replace some of the compounds from the DEL once the synthesis is completed. Hence, it is no longer sufficient to analyze it only on the level of individual molecules, but a global representation of a compound library is needed.

Here, we wish to formalize the concept of *chemical library space* (CLS)—a vector space in which residing items are entire chemical libraries. The key point here is the chemically meaningful definition of libraries as mappable objects—a generalization of standard chemical cartography. Several approaches of the representation and comparison of chemical libraries were proposed so far. For example, in the approach of Fourches *et al.*,[4] each library was represented as a similarity graph (*chemical space network*) where two nodes—individual compounds—are connected if the similarity between them is higher than a given threshold. To compare two libraries, *connectivity indices* are calculated for the corresponding graphs, allowing discrimination between similar *versus* dissimilar pairs of datasets. However, the explicit pairwise compound-to-compound similarity calculations limit the application of this approach to rather small datasets. To solve this problem, modification of the fingerprint-based similarity metrics for library comparison, avoiding calculation of the entire similarity matrix, was introduced by Miranda-Quintana *et al.*[5] Proposed extended similarity coefficients were then applied for the visualization of the similarity relationships between libraries *via chemical library networks*[6] by analogy to above-mentioned *chemical space networks*.

The aforementioned methods, however, do not intuitively explain *why* some libraries are said to be similar. Indeed, a visual pairwise inspection of compounds in the connected nodes of chemical space networks answers the question for individual molecules, but not for compound libraries. One of the methods that address this problem is a *consensus diversity plot* where library position in the CLS is defined by the pair of diversity values—(i) the median of the pairwise Tanimoto scores over intra-library compound pairs and (ii) the fraction of scaffolds retrieving 50% of the library.[7] The relative size of the collection is represented by the size of the circle representing a data point, while its color is defined by the third diversity metric—the mean of the intra-set Euclidean distance of six physicochemical properties. Such plots are easily interpretable, as each of the values in the vector has a particular chemical meaning. However, the comparison of the internal diversity of libraries instead of the similarity between them is much less informative: a library can be internally highly diverse but have a very similar chemical composition to another equally diverse library. In another library representation by a Database Fingerprint (DFP), proposed by Fernández-de Gortari *et al.*,[8] the on-bits correspond to the most frequent fragments occurring in numerous molecules from the analyzed library. Even though the DFP allows the incorporation of the main structural information of the library, it ignores finer differences between the collections that might lie in the distribution of the less frequent structural fragments or mutual occurrence and rearrangements of several fragments in different groups of compounds. There is also no possibility to include property information along with the structural one into the comparison using DFPs.

To solve the foregoing limitations of existing methods, here we introduce and test several more complex vector-based representations for compound libraries that allow us to compare numerous large collections (in our case DELs) from different perspectives and produce intuitive visualizations of the CLS. They all are based on generative topographic mapping (GTM)—a probabilistic dimensionality reduction method.[9] For each mapped item of the initial, high-dimensional descriptor space, GTM provides a vector R ("responsibility vector") rendering its fuzzy levels of assignment to the k nodes of the 2D map grid. The sum of R vectors over all members of the library provides a cumulated responsibility vector (CRV), a "baseline" representation of the library/mixture as a whole. Different refinements of this vector are introduced here:

(i) Normalized CRV (Φ), as a library-size independent library descriptor

(ii) Library-modulated CRV (Λ)—representing a library with respect to its overlap with a reference collection

(iii) Property-modulated CRV (Ω)—introducing property-centered library representation considering both chemotype and property distributions over the chemical space.

In the present article, these vectors were used to encode the previously generated 2.5k different DELs.[10] The ability of each of the vectors to accurately represent and identify DELs closest to the reference library was evaluated and compared to previous results obtained using responsibility patterns (RPs).[10] Based on the values from each of the introduced library vectors (Φ, Λ and Ω), GTM landscapes (described in detail in the Methods section) were created enabling visualization of the chemical space of a particular library from different perspectives—either from structural or property point of view and which allowed us to chemically interpret the similarity ranking results.

In more general terms, this work showcases how to exploit the flexibility of GTM technology to define inter-library similarity metrics based on different criteria—from those based on plain library overlap to scores that are fine-tuned by external information specific to each library's space zone, as captured in the herein proposed CLS vectors. Including this external information (such as the mean of calculated or measured property values) is easy and computationally efficient, because it is assigned to the "intrinsic" zones of the chemical space (the GTM nodes), *not* to the individual molecules of each library. This methodology allows one to quickly decide how much a pair of libraries *specifically* overlap within their chemical space zones characterized by desired physicochemical parameters, rather than how well they overlap "in general".

2. DATA

2.1. ChEMBL. The ChEMBL dataset (version 28) was used here as a reference library. It was downloaded and standardized in our previous work[10] according to the approach implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics at the University of Strasbourg, using the ChemAxon Standardizer.[11] This procedure 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), and generation of the major tautomer according to ChemAxon. It resulted in 1,853,565 unique ChEMBL compounds. This set is extremely diverse: for example, molecular mass spans a range between 7 (Li^+, a normorhythmic agent) and 2255 g/mol. In principle, there is

no limitation in size or complexity for molecules in DELs. In practice, however, given the peculiar constraints of the synthesis which may not work with arbitrarily complex BBs, it is clear that a part of the chemical space spanned by ChEMBL is out of the scope of any practicably achievable DEL. Hence, ChEMBL was filtered to exclude such molecules. The following filtering rules were deduced (herein tentatively named *DEL-likeness* rules), with cutoffs chosen to encompass more than 90% of all compounds in all 2497 herein considered DELs:

- $250 \leq MW \leq 750$;
- $\log P \leq 7$;
- number of H-bond acceptors ≤ 15;
- number of H-bond donors ≤ 8;
- number of rotatable bonds ≤ 15.

After filtering, 13% of ChEMBL compounds were discarded. The remaining 1,605,370 molecules were used as a reference collection in this analysis.

2.2. DNA-Encoded Libraries. 1M representative subsets for all 2497 DELs were generated in our previous work[10] with the help of the eDesigner tool.[12] This was done using commercially available BBs from eMolecules and Enamine that satisfy the Ro2[13] and eDesigner built-in DNA-compatibility filters. The enumerated compounds were standardized in the same way as the ChEMBL dataset.

3. METHODS

3.1. Generative Topographic Mapping. In chemoinformatics, each molecule can be represented as a data point defined by a vector of numerical values called descriptors. Molecules populate a chemical space, which is a high-dimensional vector space. To analyze and comprehensively visualize it, dimensionality reduction methods are needed. GTM[14−16] was the herein-used dimensionality reduction tool. It works by fitting a manifold (flexible hypersurface) into the multidimensional descriptor space populated by "frame" items, followed by the projection of the data points onto the thereupon defined 2D latent space grid.

The manifold is defined by a grid of Gaussian radial basis functions. It is fitted to the data so as to approximate the data distribution of the training set and to maximize its likelihood (*i.e.*, minimize the distance between the manifold and training data "frame" points). In more detail, the GTM algorithm training process proceeds by "bending" the manifold to pass through the densest regions of the data cloud formed by the frame set. Items are then projected from the multidimensional space onto the manifold by association to several closest grid nodes. Next, the manifold is unfolded to obtain a 2D map. The degree of association of each item (molecule or reaction, in chemoinformatics) to a node of the map is called a "responsibility". Each item is described by a responsibility vector (real number vector summing up to 1 over all nodes) that is used to define a projection of the molecule on the map. Summing up the responsibility values in each node over all molecules in the analyzed collection produces a cumulated responsibility vector (CRV) characterizing a whole library.

Different types of GTM *landscapes* can be created for the same library, where properties of the compounds projected onto each node are rendered using a color code. Three major types of landscapes were used in this study:

(1) Density landscape—created by coloring the GTM in accordance with the quantitative distribution of compounds over the nodes
(2) Library-comparative landscape—obtained by coloring the GTM by a proportion of compounds of the analyzed library in the node's overall population (populated by both analyzed and reference library molecules)
(3) Property landscape—obtained by coloring the GTM by responsibility weighted average of compound property values for each node

Using these landscapes, GTM can be applied for chemical space analysis, library comparison, or even virtual screening.[15,17]

In the present work, the first Universal GTM (UGTM)[14,17] was used for the analysis of the 2497 DELs and filtered ChEMBL28. It was built using ISIDA atom sequence counts with the length of 2−3 atoms labeled by CVFF force field types and formal charge status as descriptors.[18] Since this map was trained to predict the biological activity of molecules against 236 targets, it is suitable for the analysis of biologically relevant chemical space. It can serve not only for predictions of bioactivity but also for the analysis of large chemical libraries in the context of medicinal chemistry.[15]

3.2. Chemical Library Space. The conventional way of library analysis consists of a detailed investigation of its compound space where each compound is defined by molecular descriptors—in our case ISIDA fragment counts.[18] These fragments composed of elements of the molecule and their combinations define molecular properties. However, the structural fragment level is too detailed for characterizing the whole library. It makes little sense to build a cumulated count of all fragments seen in the members of a library because this vector loses the key information on how those fragments were initially distributed in individual compounds. In order to generalize the structural information of the library, one way would be to somehow encode the "chemotype" counts—the number of compounds of a particular "chemotype" present in a library. However, the detailed structural analysis of the large compound collection can be very computationally demanding, and the notion of "chemotype" is intrinsically vague and context-dependent.

Hence, in this work, we propose several methods of chemical library encoding derived using GTM. Since the latter preserves the topology of the initial space upon the dimensionality reduction, it is considered for the analyzed library:

(i) zones of the map are associated with predominant "chemotypes"[15,19] as implicitly defined by the highly relevant fuzzy clustering mechanism of the GTM approach
(ii) cumulated density for those zones implicitly reflect the chemotype distribution, without the need to explicitly predefine "chemotypes".

3.3. Chemical Library Encoding Methods. Several ways to use GTM responsibilities for library encoding are described in more detail below—responsibility pattern fingerprints (Γ), responsibility pattern count vectors (Γ_w), and several types of modified CRVs (Φ, Λ and Ω).

3.3.1. Responsibility Pattern Fingerprints (Γ) and Vectors (Γ_w). Due to the probabilistic nature of GTM, a position of a compound on the map is defined by a probability distribution over the nodes, which, in turn, could be encoded by a responsibility vector. Therefore, two different yet similar

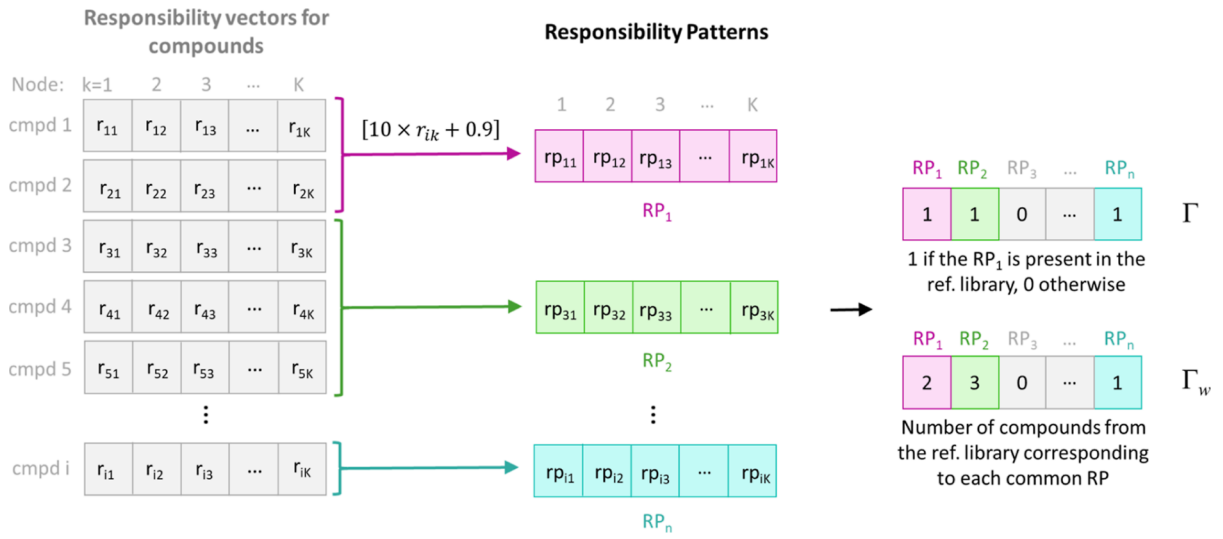


Figure 1. Summary of the RP-based library representations. The Γ values for a particular library are assigned based on the presence or absence of a certain RP in the reference library, and the Γ_w values represent the counts of reference library compounds covered by this RP.

compounds may not have exactly the same responsibility vector. However, similar compounds still are projected onto the map in a similar manner—according to a RP[20] representing discretized responsibility vector according to eq 1

$$rp_{ik} = [10 \times r_{ik} + 0.9] \quad (1)$$

where [] means truncation, rp_{ik} is the RP value for compound i in the node k, and r_{ik} is responsibility value for compound i in the node k

It follows from eq 1 that responsibility values smaller than 0.01 are reassigned to zero, and all others—to integer numbers from 1 to 10. Molecules situated close to each other in N-dimensional descriptor space and having slightly different responsibility vectors may have the same RP. These compounds usually share the same scaffold or substantial (connected or disconnected) maximum common substructure, or pharmacophore.[21] Thus, in a way, an RP could be associated with a prevalent "chemotype".

To encode a compound library using RPs, a library responsibility pattern fingerprint (Γ) and RP count vector (Γ_w) are suggested. Γ is a binary vector encoding the presence or absence of a particular reference RP in the analyzed library, and Γ_w is a vector with numerical values corresponding to the number of reference library compounds associated with each common RP present in both libraries. A schematic representation of the Γ and Γ_w calculation is given in Figure 1.

3.3.2. Normalized CRVs (Φ). A CRV = $(c_1,c_2,...,c_k)$ is the vector encoding a library by the sum of responsibility values over all molecules of the library in each node of the map, as shown in eq 2. In other words, to some degree, this vector allows the encoding of a library by the number of compounds associated with each node of the corresponding GTM plot. Thus, the CRV mathematically describes compound distribution over the 2D map and consequently over the chemical space of the library that this map visualizes. Considering that each area of the map is populated by a particular prevailing chemotype, the CRV is a crude indirect way of assessing the occurrences of different chemotypes in the library without actually defining them.

$$c_k = \sum_{i}^{N} r_{ik} \quad (2)$$

where r_{ik} is responsibility value of the molecule i in the node k

The CRV is intrinsically dependent on the size of the library it encodes. Therefore, when collections of different sizes are compared in a context in which size differences are not relevant, c_k must be normalized by library size N according to eq 3. The resulting normalized CRV (Φ) encodes relative compound distribution over the chemical space of the analyzed collection.

$$\Phi_k = \frac{c_k}{N} \quad (3)$$

3.3.3. Library-Modulated CRV (Λ). So far, the CRV and Φ consider all the chemical space zones (nodes) to be equally important in describing the library. However, some nodes may be more important—for example, the ones found to be highly populated by reference library compounds. For this purpose, the CRV of the analyzed library (a) can be modulated with respect to the compound distribution of another reference collection (r). The resulting library-modulated CRV (Λ) can be computed from the Φ of both collections, by calculating the fraction of compounds of the analyzed library in the total population of each node, as shown in eq 4. In Λ, a value $\Lambda_k = 0$ is assigned to all empty nodes in both analyzed and reference libraries, whereas for all non-empty nodes $1 \leq \Lambda_k \leq 2$ vary as a function of the fraction of compounds from the analyzed library in a given node. Nodes populated exclusively by compounds from r and a have value $\Lambda_k = 1$ and 2, respectively, whereas mixed nodes containing compounds from both libraries have values in the range $1 < \Lambda_k < 2$.

$$\Lambda_k = 1 + \frac{\Phi_k(a)}{\Phi_k(a) + \Phi_k(r)} \quad (4)$$

where Λ_k is Λ value in a given non-empty node k for analyzed library a, whereas $\Phi_k(a)$ and $\Phi_k(r)$ are normalized cumulated responsibilities in the node k for the analyzed and reference library, respectively.

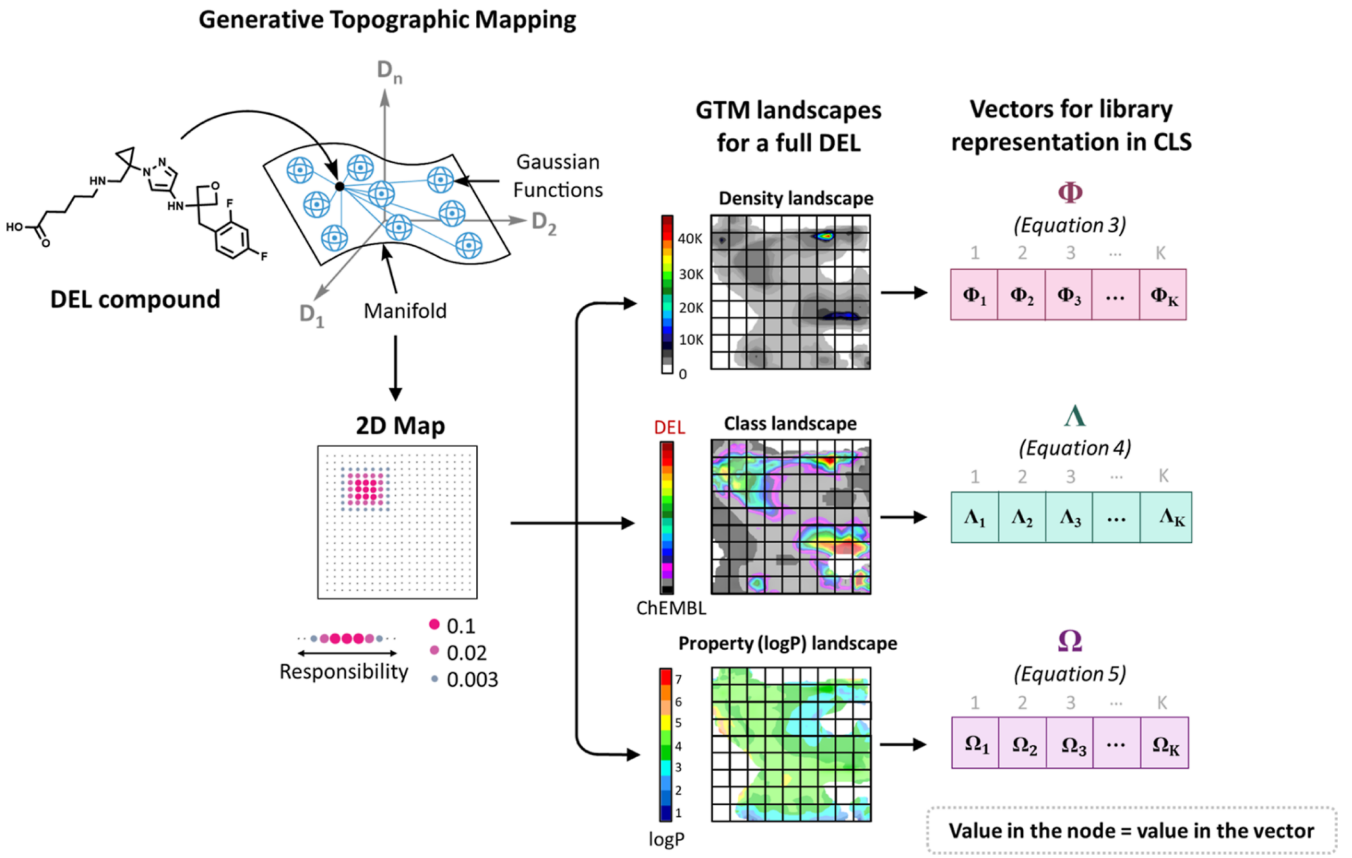


Figure 2. Scheme depicting how each of the introduced herein library encodings (Φ, Λ, and Ω) are derived from the GTM for a particular compound library.

When aiming to maximize representation and coverage of the reference collection by the analyzed library, the ideal case would be an Λ with $\Lambda_k = 0$ for the fully empty nodes and $\Lambda_k = 1.5$ (corresponding to equal representation of both reference and analyzed libraries) in all occupied ones. This "ideal" vector can thus be used as a reference in Tanimoto calculations for ranking libraries based on Λ.

3.3.4. Property-Modulated CRV (Ω). If the analysis of CLS should be performed in the context of some property or biological activity of underlying compounds for each library, the property-modulated CRV (Ω) can be used. Ω is composed of the mean property values for each node calculated according to eq 5.

$$\Omega_k = \frac{\sum_{i=1}^{N} P_i \bullet r_{ik}}{c_k} \quad (5)$$

where Ω_k is the mean property value in the node k and P_i is the property value for the compound i

Figure 2 shows a simplified scheme describing links between modified CRVs and related GTM landscapes. As soon as the compounds are projected on the map, the three types of landscapes—density, library comparative, and property landscapes—are generated, followed by preparation of related vectors Φ, Λ, and Ω using, respectively, the density, libraries ratio or mean property value in each node. Each of these vectors allows encoding a chemical library as an object in the high-dimensional CLS.

3.4. Similarity Relationships between Libraries in the CLS. To define similarity relationships between libraries in the CLS, various scores based on RP-based representation can be suggested. A score assessing the coverage of a reference library r by a candidate library a can be defined in terms of the binary Γ as the fraction of RPs of a reference library also present in a. Considering the binary nature of Γ, the coverage score is the number of on-bits common for two libraries divided by the total number of on-bits in the reference collection; see eq 6.

$$\text{Coverage}(a, r) = \frac{\sum_i \Gamma_i(a)\Gamma_i(r)}{\sum_i \Gamma_i(r)} \quad (6)$$

where the denominator simply stands for the total number of RPs encountered in the reference and $\Gamma_i(a)$ is a value (1 or 0) in the Γ of the analyzed library corresponding to the i-th RP.

However, this coverage score does not account for the number of compounds corresponding to each RP, although different RPs can be populated differently. This means that the high RP coverage does not necessarily imply high compound coverage. To solve this problem, a weighted RP coverage score can be defined as the fraction of compounds of a reference library that corresponds to the RPs present in both analyzed and reference libraries.

$$\text{wCoverage}(a, r) = \frac{\sum_i \Gamma_{wi}(r)\Gamma_i(a)}{N_r} \quad (7)$$

where $\Gamma_{wi}(r)$ is the number of compounds from the reference library r corresponding to i-th RP and N_r is the total number of compounds in the reference library r.

Notice that both coverage and weighted coverage scores were used in our previous work[10] for the comparison of virtual DEL collections with the ChEMBL database.

For the CRV-based representations (Φ, Λ, Ω), a pairwise Tanimoto coefficient is a reasonable estimation of libraries' similarity

$$Tc(a, r) = \frac{\sum_k^K v_k(a)v_k(r)}{\sum_k^K v_k{}^2(a)+\sum_k^K v_k{}^2(r) - \sum_k^K v_k(a)\cdot v_k(r)} \quad (8)$$

Here, v is a chosen CRV-based representation (v = Φ, Λ, Ω), and K is the total number of nodes.

4. RESULTS AND DISCUSSION

The herein proposed library encoding vectors Φ, Λ, Ω, Γ,and Γ_{wi} provide different views of the CLS. To investigate their usefulness, the pool of 2.5k previously generated DELs[10] was used. Three case studies were performed. First, we analyzed how proposed encodings and similarity metrics handle the comparison of a large 88 M DEL with its 1 M representative subset. The second case study addresses the selection of the "optimal" DEL for the primary screening when no or little information about the biological target is known. The goal was to identify a DEL that covers "biologically relevant" space (represented by ChEMBL) to the highest extent. For this purpose, 2.5k DELs were compared to ChEMBL (as a reference collection) in the CLS defined by Γ, Γ_w, Φ, and Λ. In the third case study, the property-focused analysis of the libraries was performed using the Ω encodings.

4.1. Representative DEL Subset *vs* Its Parent Library: A Test Study of Expected Near-Perfect Overlap. In our previous study,[10] representative sets of each of the 2.5k DELs were generated using random sampling of BBs in the eDesigner[12] tool and not the full libraries. Such a sub-library should be very similar to the entire DEL and cover virtually all of its chemical space. Therefore, overlap analysis of a representative DEL subset with respect to its parent library is a baseline case for illustrating how well each of the encodings reflects its close relationship.

For this purpose, a 3BB DEL2568 based on the aldehyde reductive amination, Migita thioether synthesis, and amine guanidinylation was selected. The coverage of the entire 88M DEL2568 by its representative subset or its similarity was calculated using each of the selected encodings (Γ, Γ_w, Φ, and Λ). Figure 3 provides a visualization of the chemical space of those two libraries. Relative compound distribution over the maps is almost identical, which backs up the claim of representation of the subset.

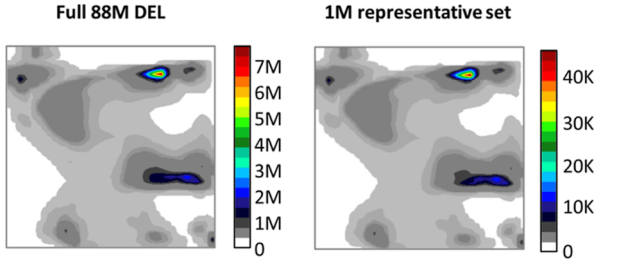


Figure 3. Density landscapes of the entire 88M DEL2568 and its 1M representative subset.

From Table 1, it appears that coverage based on Γ is very low—only 9% of RPs present in the entire DEL library are

Table 1. Coverage and Similarity of the Full DEL2568 by Its Representative Subset

CLS encoding	coverage of the full DEL2568 by the 1M subset
Γ	0.09
Γ_w	0.87
CLS encoding	Tanimoto similarity between the full DEL2568 and 1M subset
Φ	0.99
Λ	0.98

covered by the 1M representative set. However, Γ_w coverage shows that those 9% of RPs correspond to 87% of molecules, which means that the subset lacks very rare (but numerous) RPs, all while covering "mainstream" chemotypes from the full collection. It is interesting to witness a combinatorial library (sharing a common "scaffold" defined by the underlying chemistry) concentrating 87% of its members into 9% of the spanned chemical space. This is not unexpected—combinations of relatively "exotic" and rare BBs result in "exotic" but rare products.

The similarity between those two collections was also calculated using CRV-based representations—Φ and Λ. In the latter case, the Λ vector of the 1M subset was created by calculating the ratio of molecules from the representative subset with respect to the reference (full 88M collection) in each node of the map. It was then compared to the "ideal" Λ where each node occupied by the reference 88M library has a value Λ_k = 1.5, which corresponds to the perfect representation of the full library by the subset (see details in the Methods section). Tanimoto coefficients calculated for CRV-based representations are given in Table 1. Those values being close to the maximum illustrate expected (and observed in Figure 3) high similarity between compound distribution in the chemical spaces of those libraries.

Both CRV-based representations provide close to the maximum similarity values between the library and its representative subset, as expected. RP-based representations, on the other hand, provide a stricter comparison with an accent on the missing reference RPs (chemotypes) in the analyzed library. This example demonstrates the importance of using both the Γ- and Γ_w-based coverage scores. While the first one shows how many "chemotypes" are covered, the second one puts this number into the perspective of their compound population and provides a compound-weighted coverage of the chemical space.

4.2. ChEMBL *vs* DEL Comparison in the CLS Defined by Different GTM-based Encodings. As in our previous work,[10] here we focused on the case of primary screening where the selected DEL needs to cover the biologically relevant chemical space to the highest extent. Technically, such a task consists of ranking the 2.5k DELs by their similarity (or coverage) to a reference collection—here, the ChEMBL database.

4.2.1. Library Comparison by Responsibility Distribution. Coverage and Tanimoto similarity coefficients for each of the 2.5k DELs were calculated with respect to the ChEMBL library using each of the encodings (Γ, Γ_w, Φ, and Λ). The results are combined in Figure 4. Two libraries—DEL2568 and DEL271 having the highest and the lowest weighted ChEMBL coverage

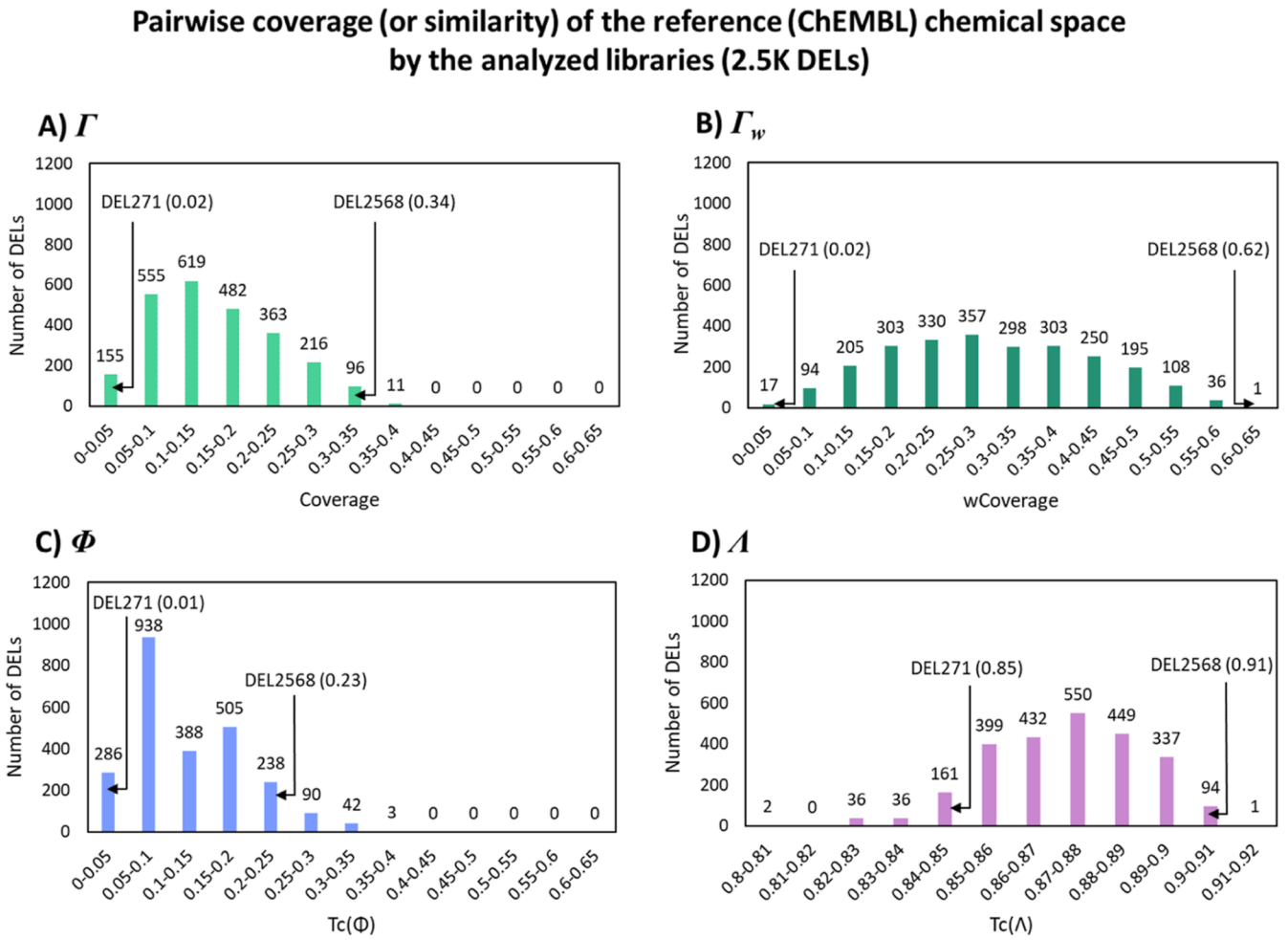


Figure 4. Pairwise comparison of 2.5k DELs with ChEMBL using different representations and metrics: distribution of ChEMBL coverage scores calculated using Γ (A) and Γ_w (B), and distribution of Tc between ChEMBL and each DEL calculated using Φ (C) and Λ (D).

based on Γ_w—were selected as points of reference, to trace their scoring with other representations. Both Γ-based coverage (Figure 4A) and Φ-based Tc (Figure 4C) adopt values within a similar and rather low value range: from 0.01 to 0.4. This highlights that DEL compound distribution is quite different from that of ChEMBL, and the likelihood of finding the ChEMBL RPs in DELs is rather low. However, the Γ_w-based coverage shows that those RPs that are covered by DELs in fact correspond to the prevailing compound population of ChEMBL because observed values of coverage almost doubled with respect to Γ-based coverage (Figure 4B). In all three cases, the two "marker" libraries, nevertheless, keep their relative rank: DEL2568 is always ranked in the top 5−10% of libraries and DEL271—in the last 10−15%. As expected, tuning the overlap criterion by means of the usage of different CLS vectors should never override the fundamental "core" library similarity, distinguishing between libraries containing closely related molecules from those which do not.

In the case of Λ-based similarity, the Tc values are spread within a narrow range: from 0.8 to 0.92 (Figure 4D). The Λ-based similarity spectrum is intrinsically different from those calculated using other encodings. Since vectors for all libraries are modulated with the CRV of the same reference collection, the similarity value between two Λ is always higher than that in the case of Φ, for example. However, the position of DEL2568 and DEL271 in Figure 4D is similar to the other three cases. Thus, even though being shifted toward higher values, similarity distribution in the CLS defined by Λ follows the same trends as in other library spaces.

For further analysis of the similarity relationships in the four proposed representations of CLS, all DELs were ranked with respect to the coverage of (or similarity to) ChEMBL. To simplify the analysis, here we analyze only five DELs: ranked the first, 50th, 100th, 1000th, and 2497th with respect to ChEMBL. For each of these five DELs, a density landscape showing compound distribution in the chemical space of the library was created (see Figure 5). This figure shows that each of the representations ranks libraries differently—none of the libraries were selected as the best one by more than one representation. However, DELs having the same rank in different spaces (landscapes forming columns in Figure 5) still have very similar compound distribution over the map. Failure to consensually score one DEL as the best match for ChEMBL, in any CLS, is due to the fact that there are several DELs that might claim this title, and no single one is undoubtedly outstanding in terms of sharing related chemotypes with ChEMBL. Looking at the problem through the prism of multiple CLS definitions is evidencing this important aspect, that is, allowing for more flexibility in experimental setups. In this scenario, there is no particular reason to pick either of the DELs of column no 1 in Figure 5—a case in which extraneous parameters (availability, facility of synthesis, and cost) may be

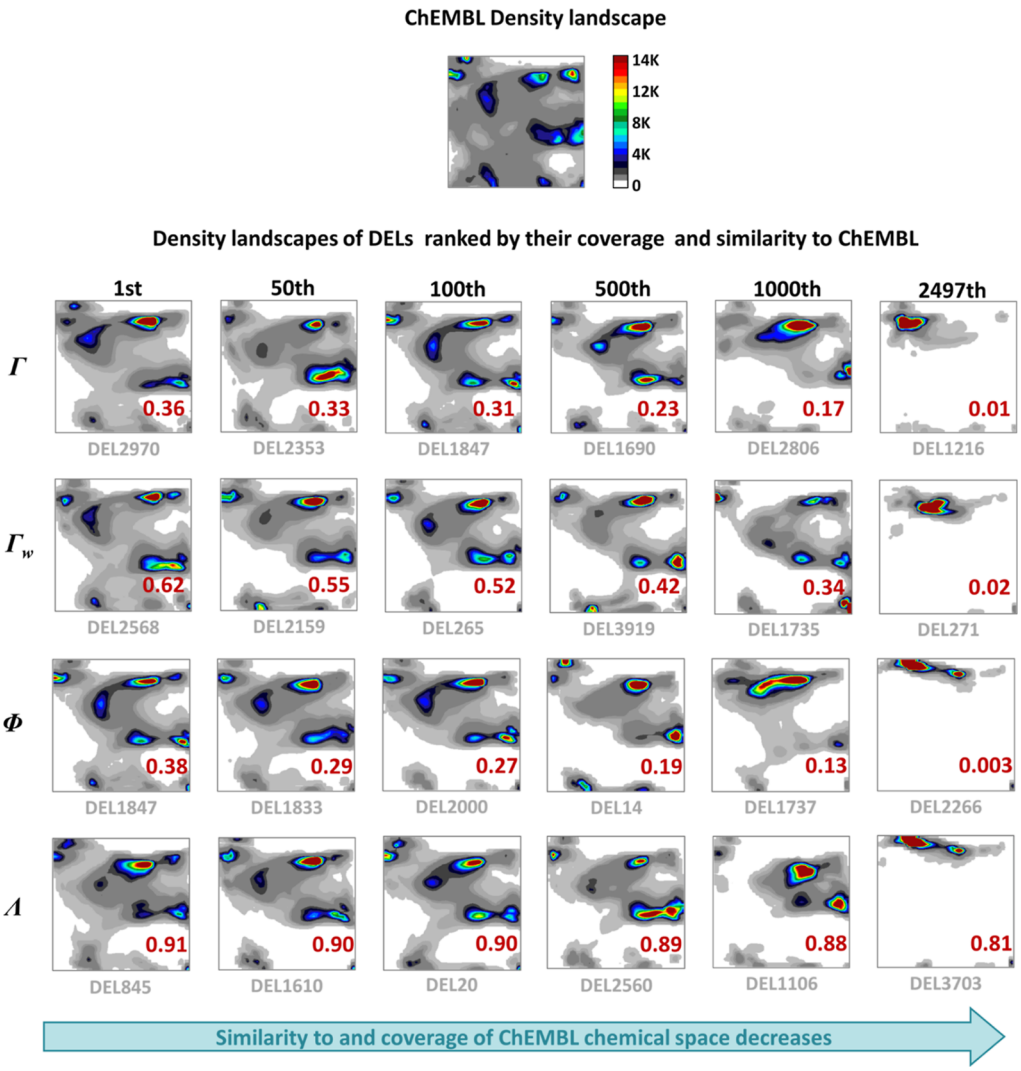


Figure 5. Density GTM landscapes of ChEMBL28 and selected DELs ranging from the most similar to the least similar to ChEMBL. DELs were selected and ranked either by coverage scores (in the case of Γ and Γ_w) or Tanimoto similarity coefficients (in the case of Φ and Λ). Values of either coverage or Tc are provided in red on each landscape. For all landscapes, the same color scale corresponding to the density distribution of ChEMBL was used.

applied by the user to select either of these. Should a consensual winner emerge from this analysis, selecting it at higher costs over the others may make sense. Practically, however, visual inspection shows that the first few hundred DELs have similar density landscapes to the top-ranked landscapes corresponding to the 100th or even 500th-ranked library still match the landscapes in column 1 quite well. Finally, yet importantly, within the top 100 DELs chosen by each of the encodings, there are 32 DELs common to all four encodings; within the top 500, this value rises to 273, and for the top 1000 DELs, it reaches 713, which shows how well the ranking by coverage or Tc based on four encodings correspond to each other. For more details, see Figure S1 of Supporting Information.

Even though each of the analyzed representations offers a different DEL as the closest to ChEMBL (DEL2970, DEL2568, DEL1847, and DEL845), they all appear to be quite similar. Interestingly, all these libraries are three-cycled DELs that were designed exclusively based on robust coupling reactions—aldehyde reductive amination (all four libraries), Ullmann-type *N*-aryl coupling (DEL2970 and DEL845), Migita thioether synthesis from thiophenols and arylbromides (DEL1847 and DEL2568), and carboxylic acid/amine condensation (DEL1847 and DEL845) (see Figure S2 of Supporting Information). The size of the full DELs is also very similar for those four libraries—slightly above 80M compounds. The reason for the high diversity of those collections and thus high coverage of (and similarity to) ChEMBL is due to the abundance and diversity of the purchasable BBs required for those reactions—amines, aldehydes, arylhalides, and carboxylic acids.[10,22]

Libraries with the lowest rank—DEL1216, DEL271, DEL2266, and DEL3703—also have some design features in common. Their full size is much lower (between 1M and 5M), and they all have at least two heterocyclization steps in their design—aminothiazole and Larock indole synthesis were combined to form DEL1216, imidazole and Larock indole synthesis were used in DEL271 generation, and three heterocyclization steps (oxadiazole, triazole, and aminothiazole synthesis) were used both in DEL2266 and DEL3703 (see Figure S3 of Supporting Information). As is visible from Figure 5, those collections have one (maximum two) density peak,

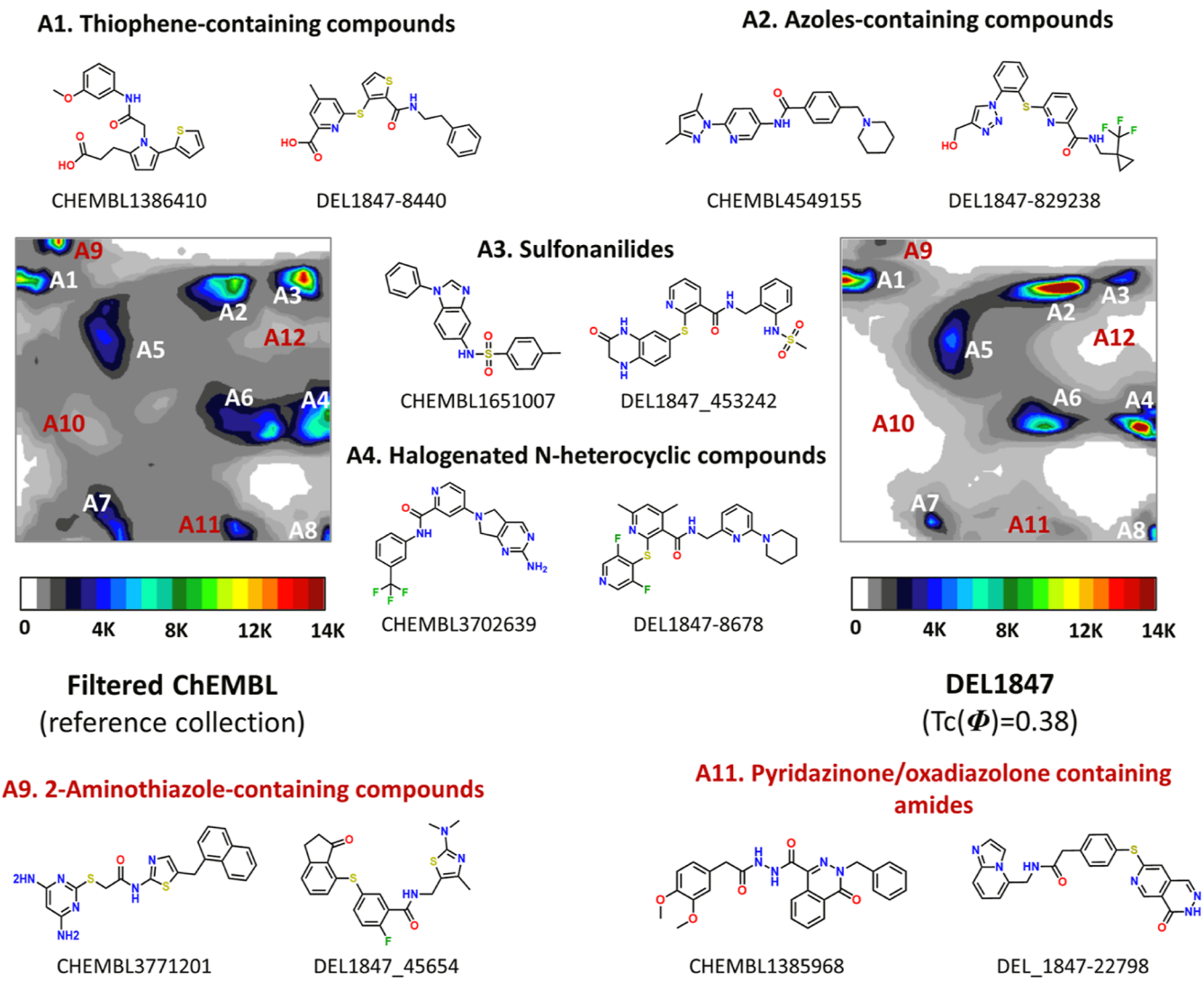


Figure 6. Interpretation of the similarity between ChEMBL and DEL1847 *via* structural analysis of the density landscapes of those libraries. Areas A1−A8 (labeled in white) correspond to the peaks of high density in ChEMBL space that were reproduced in DEL1847. Areas A9−A12 (labeled in red) represent mismatched zones.

which means that their diversity is much lower, and those DELs can be considered as focused libraries containing very similar compounds. This is explainable by the fact that employing two heterocyclization steps in DEL synthesis means that all compounds possess at least two identical heterocycles—a consequently large scaffold—with diversity being introduced only *via* their "ornaments", by contrast to, say, an amide formation in which everything but the −C(═O)NH− moiety is variable.

The use of only heterocyclizations is convenient for "focused" DEL synthesis, as the common scaffold generated by the reaction represents a common signature of all library members, which vary in terms of scaffold substituents only.[23] This provides an excellent library for extracting structure−activity relations and fine-tuning lead molecules, provided, of course, that the focus around the chosen heterocyclic core matches the actual chemical space zone favored by the target. However, if the goal is to produce general-purpose DELs, it is a safer option to use building-block-rich coupling reactions instead because abundant BB classes exist. Many BBs already contain necessary heterocyclic moieties,[24] albeit not necessarily connected to each other in a same way as they would be linked up in a heterocyclization synthesis-based DEL. Another option might be to use only one heterocyclization step combined with two coupling synthetic cycles. In this way, the diversity coming from coupling reactions can partially compensate for the presence of the same heterocycle in each molecule. An example of such design is DEL2806 (1000th library by Γ)—it combines imidazole synthesis with guanidine group formation from amines and Ullmann-type *N*-aryl coupling. All other DELs featuring from 1st to 500th in Figure 5 are based only on coupling reactions.

4.2.2. In-Depth Analysis and Interpretability of Library Overlap. Overlap scores are useful for the rapid processing and ranking of large sets of candidate libraries, but a real understanding of overlap must go down to individual compound structure levels. The strength of this protocol is that the mapping used to define CLS vectors can implicitly support this approach. To illustrate that, the density landscape for DEL1847 that is the closest to ChEMBL according to Φ ranking was compared to the density landscape of ChEMBL (Figure 6). DEL1847 is a three-step library based on aldehyde reductive amination with the NH_2 group of the headpiece (2652 aldehydes), followed by the condensation of the same

amino-group with 21 bifunctional carboxylic acids containing thiol group that on the third cycle reacts with 1630 arylbromides to form thioether bonds. The total size of the library is around 90M.

In Figure 6, most of the density peaks of ChEMBL (A1−A8) were reproduced in DEL1847. These areas contribute to the similarity of those two libraries and make DEL1847 the most highly scored by the Tanimoto coefficient (Tc = 0.38) calculated based on Φ. Indeed, areas A1−A4 are covered by both libraries, containing molecules of similar structural features, even though DEL1847 compounds also have thioether and amide groups in their structures. Nevertheless, this similarity value is far from perfect, which can be explained by mismatched density peaks between ChEMBL and DEL1847. Namely, areas A9 and A11 are heavily populated in the ChEMBL landscape, but rather moderately occupied in DEL1847. The former area is populated by 2-aminothiazole-containing compounds and is expectedly underrepresented in DEL1847, as only 14 BBs used for its enumeration contain this structural moiety (0.3% of all BBs). The same applies to area A11, which is highly populated by pyridazinone/oxadiazolone-containing amides in ChEMBL and underpopulated in the case of DEL1847. Regions A10, A12 in ChEMBL are empty in DEL1847. This is because these areas are populated by complex natural products,[10] and thus cannot be reproduced by herein considered DELs.

The same analysis was performed for the most dissimilar one to the ChEMBL library by Φ—DEL2266. This library is based on three heterocyclization reactions—oxadiazole, triazole, and aminothiazole synthesis—that provide 1.3M compounds in total. As a result, each compound of the library contains the same three cycles, which makes this library structurally highly focused. However, there are no molecules in filtered ChEMBL28 of similar chemotypes. In Figure 7, highly populated areas A1 and A2 in the DEL2266 landscape are almost empty on the ChEMBL map, and the two libraries almost do not overlap at all, which explains close to zero similarity between them.

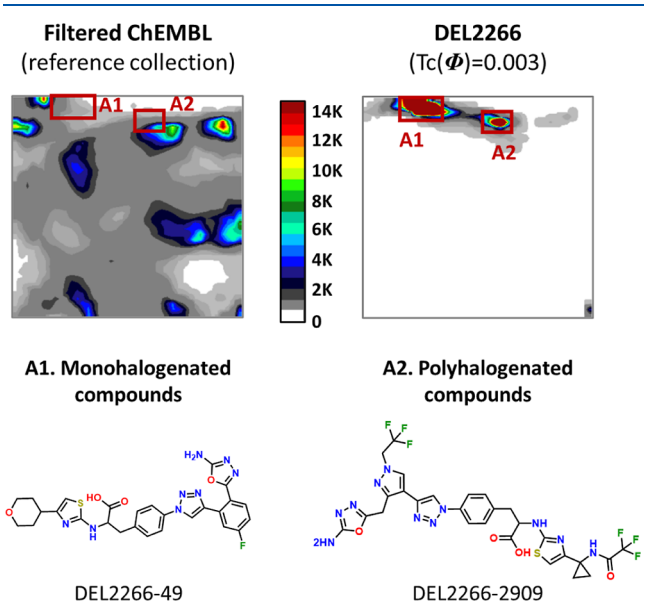


Figure 7. Interpretation of the similarity between ChEMBL and DEL2266 *via* structural analysis of the density landscapes of these libraries.

Thus, by analyzing density landscapes for the selected pairs of libraries, it is possible to explain the similarity behavior in the CLS defined by Φ. The interpretation of the CLS defined by Λ can be performed by analyzing pairwise comparative landscapes featuring reference collection against each of the analyzed libraries.

4.2.3. Property-Sensitive Library Comparison. A conventional way to analyze compound collections in terms of a particular physicochemical property is to build a frequency plot (histogram) showing the distribution of this property for all library molecules.[25−28] This approach though has several drawbacks. First of all, there is a complete disconnection of such plots from the chemotype composition of the analyzed collection. Figure 8 shows that both libraries closest and farthest to ChEMBL according to Γ_w ranking (DEL2568 and DEL271, respectively) have a very similar distribution of log *P* values, even though they strongly diverge in terms of composition. Moreover, compounds with a given property value (*e.g.*, log *P* = 4) may be spread all over the map—they do not have to be similar simply because they share the same property value (Figure 8 on the right).

By contrast, property-modulated Ω has two key advantages: being focused on specific chemical space zones populated by similar chemotypes, it does account for the chemistry "behind" the property values. The second key feature is that property-related information is provided *via* GTM property landscapes, thus it is directly associated with chemical space zones. In this way, Ω representation allows for dual libraries' analysis and comparison where the most similar to the reference library collection simultaneously demonstrates both chemotype and property similarity.

To further illustrate the advantages of Ω over the property histograms, the DELs most similar to ChEMBL were selected and compared using both approaches. First, each classical bar chart for H-bond acceptor count was encoded by a n-component vector, whose length corresponded to the number of bars in the property histogram. Then, based on these vectors, Tanimoto coefficients were calculated between each DEL and ChEMBL, and the most similar DEL2189 was selected (see Figure 9A) with Tc = 0.95. The same was done by calculating the Tanimoto coefficient between each DEL and ChEMBL using the respective Ω, which led to the selection of DEL630 as the most similar one (Figure 9C) with Tc = 0.78. The Tc values for both DEL2189 and DEL630 calculated either based on the Ω or H-bond acceptor counts distribution vectors with respect to the filtered ChEMBL database are given in Table 2.

From Figure 9 it is visible that even though having similar global property distributions (illustrated in histograms), the local distribution of H-bond acceptor counts in each area of the chemical space of DEL2189 (Figure 9A) is dissimilar compared to the ChEMBL property landscape (Figure 9B)— there are almost no zones containing compounds with more than eight hydrogen bond acceptor atoms on the DEL2189 landscape. Moreover, there are lots of ChEMBL areas that are empty on the DEL2189 landscape, thus the chemotype similarity of this library to ChEMBL is low (Tc(Φ) = 0.13). In contrast, DEL630 (Figure 9C) selected as the most similar to ChEMBL using HAC-Ω representation has a significantly larger colored surface which means higher chemotype similarity to ChEMBL (Tc(Φ) = 0.34). Furthermore, the local property distribution in this collection is much closer to ChEMBL than that in DEL2189. Indeed, there are many areas

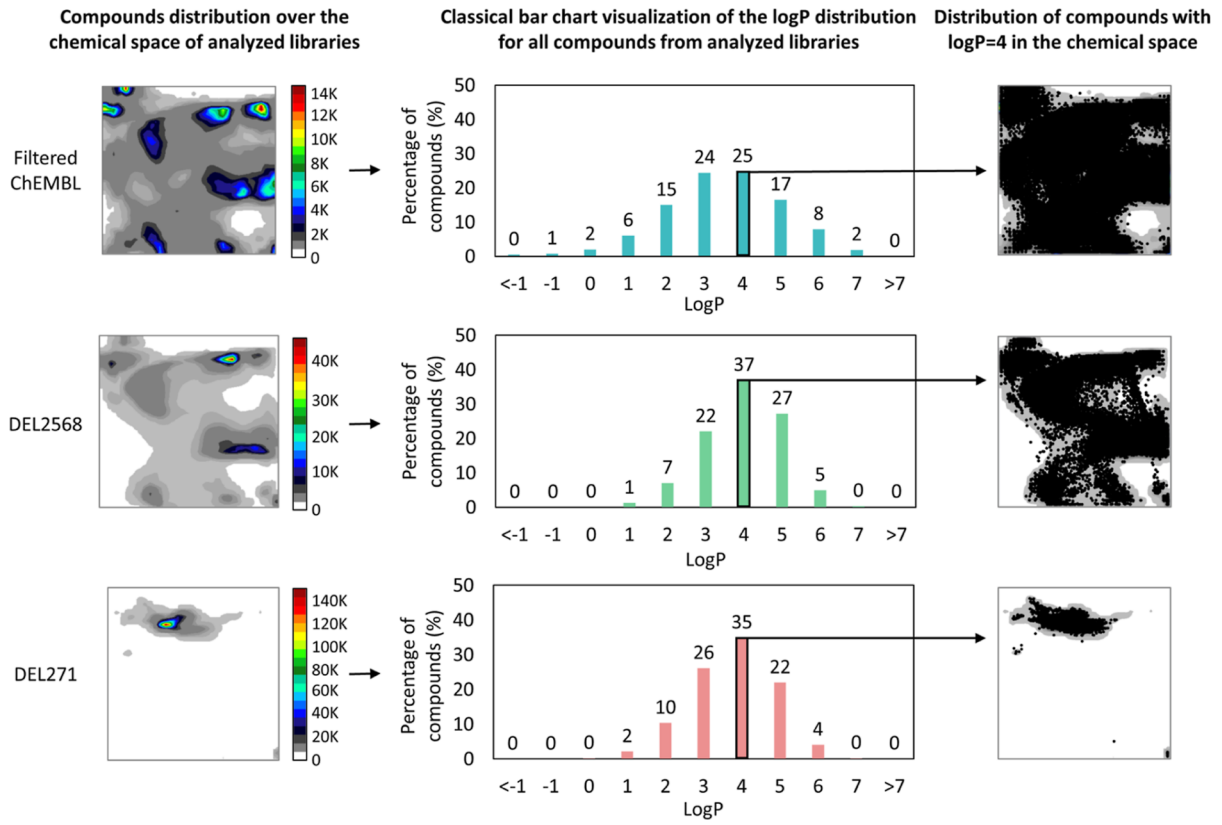


Figure 8. (Left) Density landscapes of filtered ChEMBL, DEL2568, and DEL271; (center) classical bar chart visualization of calculated log *P* distribution for all compounds from analyzed libraries; (right) compounds with log *P* = 4 (black dots) projected on the corresponding density landscapes.

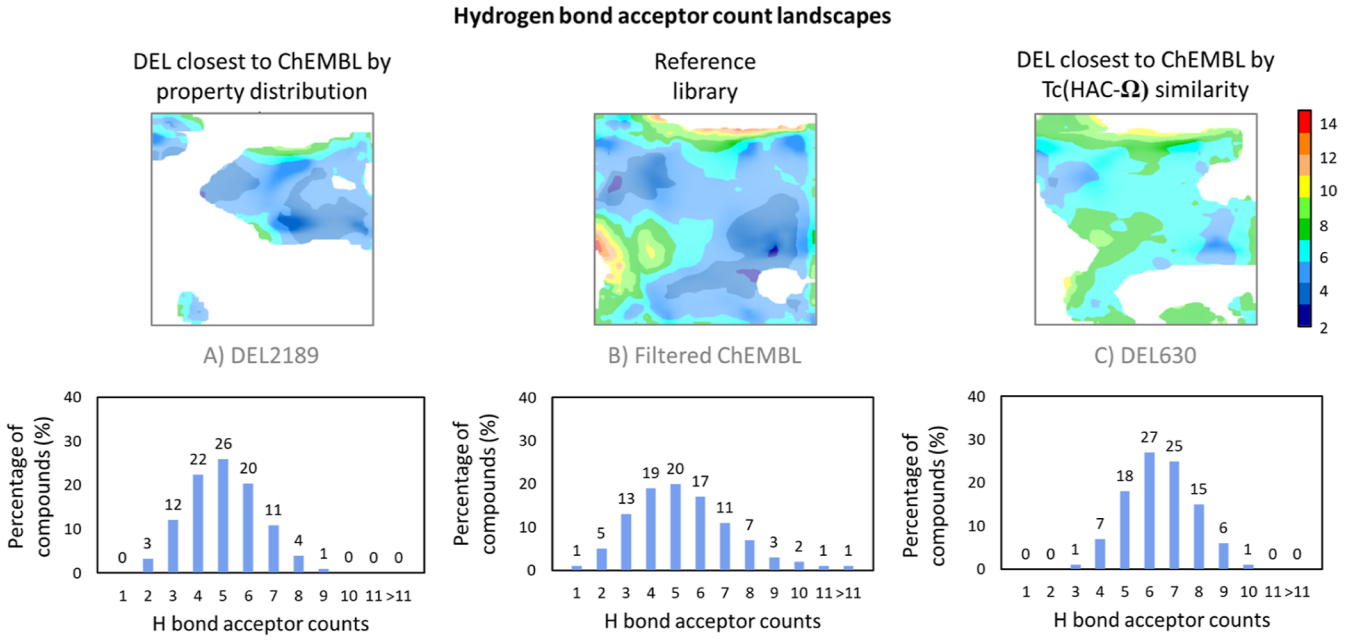


Figure 9. Hydrogen bond acceptor count (HAC) landscapes for (A) DEL2189 (selected by property distribution similarity), (B) reference library-filtered ChEMBL28, and (C) DEL630 [selected by Tc(HAC-Ω) similarity].

colored in the same way in both ChEMBL and DEL630 collections, which means that the average number of H-bond acceptors in compounds populating these zones is very close.

Thus, Ω encoding allows us to take into consideration both property and chemotype distribution in the chemical space of analyzed libraries. Different Ω can be created using any measured or calculated property if it is provided for every compound in analyzed libraries. Figure S4 renders the distribution of the similarity of DELs with respect to ChEMBL in six Ω-encoded CLS: MW, log *P*, H-bond acceptors and

Table 2. Tanimoto Values for DEL2189 and DEL630 Calculated Either Using HAC-Ω or H-Bond Acceptor Count Distribution Vectors with Respect to the Filtered ChEMBL28 Database

	Tc(HAC-Ω)	Tc(property distribution)
DEL2189	0.34	0.95
DEL630	0.78	0.67

donors, number of rotatable bonds, and quantitative estimate of drug-likeness (QED score). Using these values libraries can be ranked according to their property-focused similarity to ChEMBL. As an example, in Figure 10 six QED landscapes of DELs ranging from the most similar to the least similar to ChEMBL in the CLS defined by QED-Ω are provided. As we go from the first to the last DEL, there is a decrease in the similarity between each of their QED landscapes and the QED map of ChEMBL. The top-ranked collection—DEL45 is based on only two reaction steps (aldehyde reductive amination followed by imidazole synthesis reaction) and thus expectedly contains a lot of drug-like compounds (97% of the whole library). Thus, the QED values for this library are also higher than for molecules enumerated *via* a combination of three BBs in three cycle DELs, which we can see on the landscapes. Figure 10 also shows that there are a lot of areas on the ChEMBL and DEL45 QED landscapes that are colored in the same way. This means, that DEL45 is reproducing not only global but also local QED distribution observed in the ChEMBL chemical space. The Tanimoto coefficient value calculated in the Φ-based CLS (Tc = 0.25, DEL45 is 167th most similar to ChEMBL by Φ among 2497 DELs in total) and visual similarity between the density landscapes of those libraries prove that QED-modulated Ω encodes not only global and local property distribution but also chemotype distribution for the analyzed libraries.

5. CONCLUSIONS

In this work, we reported the development of several types of vector-based encodings for characterizing libraries of various sizes and compositions as a function of the relative distribution of molecules in the GTM-based chemical space. These representations constitute a new way of the analysis of combinatorial mixtures, such as DELs, that should be considered not only as an ensemble of compounds, but also as unified entities—mixtures whose composition cannot be easily changed once synthesized. Of course, the methodology generally applies in contexts where any library—cherry-pickable or not—needs to be regarded as a stand-alone entity, rather than a collection of individual molecules. With the encodings introduced here, it becomes possible to clearly define CLS where each collection is considered as a data point. Classical chemoinformatics allows for the management of a portfolio of *compounds* forming a core library (comparison to other compound sets, directed enrichment in new compounds, focused subset extraction for screening, *etc.*), whereas this methodology enables the management of a portfolio of *libraries* (selection of the best suited one for a screening campaign, enrichment with novel libraries—overlapping or not, *etc.*).

From the example of ChEMBL *vs* DEL comparison, it was shown that all proposed CLS representations—responsibility pattern fingerprints (Γ), responsibility count vectors (Γ_w), normalized CRVs (Φ), library-modulated CRVs (Λ), and property-modulated CRVs (Ω)—are able to efficiently encode key information about the "chemotype" distribution of analyzed libraries, where "chemotypes" are implicitly defined by the intrinsic neighborhood compliance of GTMs. "chemotypes", in this sense, may be common scaffolds including or not common key "ornaments", common topological pharmacophores, or more loosely defined compound clusters of molecules with a specific global charge or outstanding size, *etc.* Similarity relationships in all five CLSs seem reasonable and chemically meaningful and allow adequate sorting of DELs

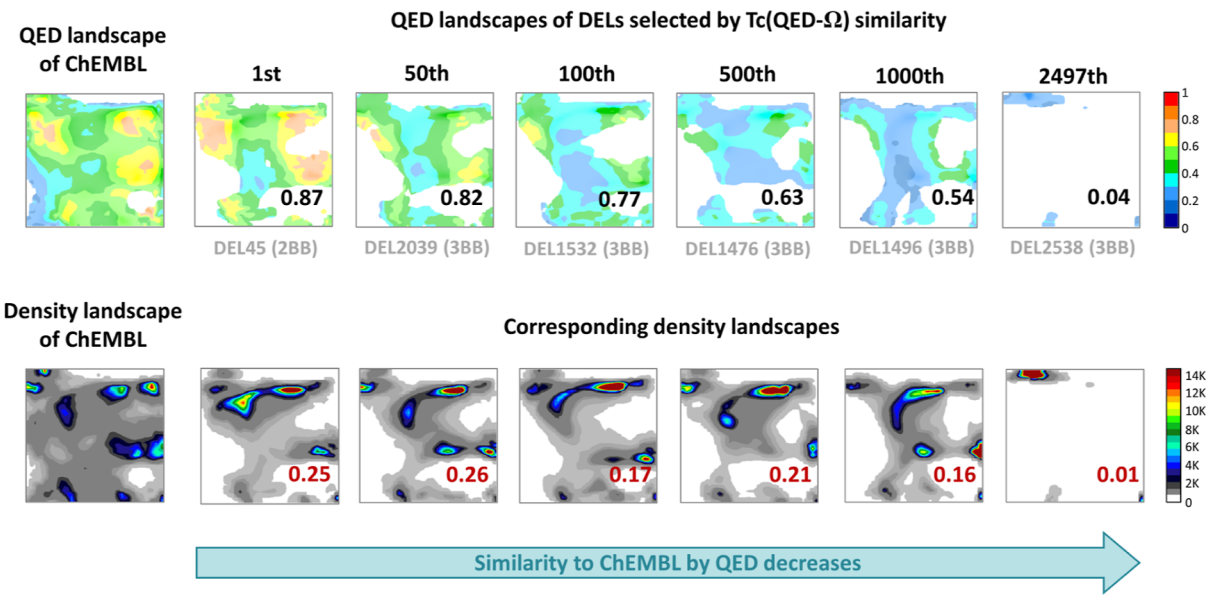


Figure 10. First row: On the left: QED landscape of filtered ChEMBL28. On the right: QED landscapes of DELs ranging from the most similar to the least similar to ChEMBL sorted by their Tanimoto coefficients calculated based on their QED-Ω with respect to ChEMBL (in black). Second row: On the left: density landscape of the filtered ChEMBL. On the right: corresponding density landscapes for selected DELs with their Φ similarity values with respect to ChEMBL (in red).

with respect to their similarity to ChEMBL. Therefore, any of the proposed representations can be used for selecting an optimal DEL for a particular task if the reference collection can be defined. Here, ChEMBL was used to represent the drug-relevant chemical space, and it was assumed that the ultimate goal in general diversity library design is mimicking the chemical space covered by it. This is of course debatable—in real applications, experts may define reference libraries based on much stricter and project-specific criteria. The present work outlines a novel methodology for library selection and comparison, which was shown to be senseful in all respects concerning the analysis of herein considered DELs, but must yet be proven useful in prospective library design—a goal unfortunately way beyond the resources of many academic research teams.

To analyze libraries with respect to the featured chemotypes without paying attention to their population the best choice would be Γ. If the population of the matched chemotypes in only one of the libraries (reference collection) is important—the coverage score based on the Γ_w should be used, thereby ensuring that the candidate library matches the often-seen patterns in the reference collection, and not its atypical "singletons". In case the compound distribution over the chemical space of all analyzed collections is important, CLS should be defined by the Φ, whereas Tanimoto similarity should be used for library ranking. This strategy can also be used in order to select a library that maximally reproduces compound distribution from the chemical space of the reference collection (*e.g.*, selection of the optimal representative subset). Λ-based encoding is particularly useful when one wants to compare a coverage of a reference dataset by some other libraries. In this case, each library is encoded considering its relative compounds distribution with respect to the reference collection, so a special accent is placed on the differences between the relative proportion of compounds coming from analyzed and reference libraries without taking into consideration the absolute popularity of each node. Moreover, in case the accent of the analysis is placed on the particular calculated or measured property, Ω can be used to encode libraries with respect to both chemotype and property distribution in the chemical spaces of these collections. In contrast to classical property histograms that describe the global distribution of the property values among compounds of the whole library, Ω encodes local property distribution among compounds belonging to different chemotypes and populating particular areas of the chemical space.

The interpretability of the proposed vectors merits a special mention here. Being GTM-based, Φ, Ω, and Λ can be visualized as compound density, property, or comparative landscapes for each library on a separate plot. By analyzing landscapes of the selected pairs of libraries, the similarity behavior in particular CLS can be investigated and interpreted. For example, in the case of Φ-defined CLS, by comparing the highest peaks on the density landscapes of two libraries it is easy to identify which common chemotypes positively contributed to the similarity, and which mismatched areas of the chemical space decreased the Tanimoto value.

Now, when the performance of the proposed encodings and the similarity behavior of libraries (objects) in corresponding CLS are analyzed and described, it should be last but not least noted that this CLS may also be visualized, like any "classical" chemical space. In perspective, the meta-GTM approach[29] is perfectly suited for the dimensionality reduction and visualization of CLS.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00520.

Venn diagrams comparing several DELs, density landscapes of selected DELs, and distributions of some physicochemical parameters of selected DELs (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Alexandre Varnek − *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg 67081, France;* orcid.org/0000-0003-1886-925X; Phone: +33 368851560; Email: varnek@unistra.fr

Authors

Regina Pikalyova − *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg 67081, France*

Yuliana Zabolotna − *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg 67081, France*

Dragos Horvath − *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg 67081, France;* orcid.org/0000-0003-0173-5714

Gilles Marcou − *Laboratory of Chemoinformatics, University of Strasbourg, Strasbourg 67081, France;* orcid.org/0000-0003-1676-6708

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00520

Funding

R.P.—Bourse de l'Ecole Doctorale des Sciences Chimiques ED222, Université de Strasbourg.

Notes

The authors declare no competing financial interest.

The data used in this work are available in the public domain resources: biologically relevant compounds from ChEMBL[30] (version 28)—https://www.ebi.ac.uk/chembl/, eMolecules[31] BBs that were used for DEL generation using eDesigner[12] are partially available on the website https://www.emolecules.com/products/building-blocks, and Enamine[32] BBs are available on the website https://enamine.net/building-blocks.

■ ACKNOWLEDGMENTS

The authors would like to thank eMolecules Inc.[31] for providing the collection of commercially available BBs that were used for the generation of DELs analyzed in this work.

■ REFERENCES

(1) Czarnik, A. W. Encoding methods for combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1997**, *1*, 60−66.

(2) Brenner, S.; Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5381−5383.

(3) Franzini, R. M.; Neri, D.; Scheuermann, J. DNA-encoded chemical libraries: advancing beyond conventional small-molecule libraries. *Acc. Chem. Res.* **2014**, *47*, 1247−1255.

(4) Fourches, D.; Tropsha, A. Using graph indices for the analysis and comparison of chemical datasets. *Mol. Inf.* **2013**, *32*, 827−842.

(5) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended similarity indices: the benefits of comparing more than two

objects simultaneously. Part 1: Theory and characteristics. *J. Cheminf.* **2021**, *13*, 32.
(6) Dunn, T. B.; Seabra, G. M.; Kim, T. D.; Juárez-Mercado, K. E.; Li, C.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Diversity and Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.* **2021**, *62*, 2186−2201.
(7) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus diversity plots: a global diversity analysis of chemical libraries. *J. Cheminf.* **2016**, *8*, 63.
(8) Fernández-de Gortari, E.; García-Jacas, C. R.; Martinez-Mayorga, K.; Medina-Franco, J. L. Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminf.* **2017**, *9*, 9.
(9) Bishop, C. M.; Svensen, M.; Williams, C. K. I. GTM: The generative topographic mapping. *Neural Comput.* **1998**, *10*, 215−234.
(10) Pikalyova, R.; Zabolotna, Y.; Volochnyuk, D. M.; Horvath, D.; Marcou, G.; Varnek, A. Exploration of the Chemical Space of DNA-encoded Libraries. *Mol. Inf.* **2022**, *41*, 2100289.
(11) ChemaAxon. *JChem*, version 20.8.3; ChemAxon Ltd: Budapest, Hungary, 2020.
(12) Martin, A.; Nicolaou, C. A.; Toledo, M. A. Navigating the DNA encoded libraries chemical space. *Commun. Chem.* **2020**, *3*, 127−129.
(13) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discovery Today* **2015**, *20*, 11−17.
(14) 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*, 1087−1108.
(15) 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.
(16) Lin, A. Cartographie topographique générative: un outil puissant pour la visualisation, l'analyse et la modélisation de données chimiques volumineuses. Ph.D. Thesis, Université de Strasbourg: Strasbourg, 2019.
(17) 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.* **2018**, *59*, 564−572.
(18) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.
(19) Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping in drug design. *Drug Discovery Today: Technol.* **2019**, *32−33*, 99−107.
(20) 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*, 1438−1454.
(21) Kayastha, S.; Kunimoto, R.; Horvath, D.; Varnek, A.; Bajorath, J. From bird's eye views to molecular communities: two-layered visualization of structure−activity relationships in large compound data sets. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 961−977.
(22) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath, D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek, A. A close-up look at the chemical space of commercially available building blocks for medicinal chemistry. *J. Chem. Inf. Model.* **2021**, *62*, 2171−2185.
(23) Dickson, P.; Kodadek, T. Chemical composition of DNA-encoded libraries, past present and future. *Org. Biomol. Chem.* **2019**, *17*, 4676−4688.
(24) Oksiuta, O. V.; Pashenko, A. E.; Smalii, R. V.; Volochnyuk, D. M.; Ryabukhin, S. V. Heterocyclization vs Coupling Reactions: A DNA-Encoded Libraries Case. *Zh. Org. Farm. Khim.* **2023**, *21*, 3−19.
(25) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.
(26) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug-and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470−479.
(27) Lucas, X.; Gruning, B. A.; Bleher, S.; Gunther, S. The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* **2015**, *55*, 915−924.
(28) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K.-C. Assessment of chemical libraries for their druggability. *Comput. Biol. Chem.* **2005**, *29*, 55−67.
(29) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J. Chem. Inf. Model.* **2015**, *55*, 84−94.
(30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(31) eMolecules Inc. https://www.emolecules.com/.
(32) Enamine Ltd. https://enamine.net/.