



HAL
open science

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

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

► **To cite this version:**

Regina Pikalyova, Yuliana Zabolotna, Dragos Horvath, Gilles Marcou, Alexandre Varnek. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *Journal of Chemical Information and Modeling*, 2023, 63 (13), pp.4042-4055. 10.1021/acs.jcim.3c00520 . hal-04243839

HAL Id: hal-04243839

<https://hal.science/hal-04243839v1>

Submitted on 16 Oct 2023

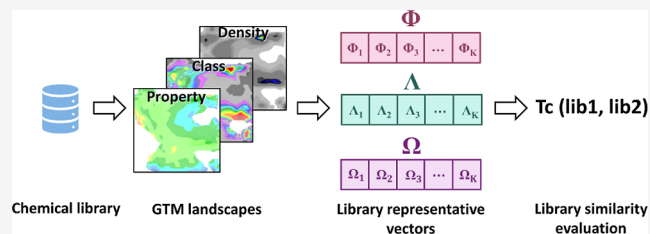
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

4 **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 stand-alone 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

22 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.

45 With time, combinatorial chemistry has advanced to the point that it is now possible to simultaneously synthesize a

47 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 64

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

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

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

126 To solve the foregoing limitations of existing methods, here
127 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 assign-
ment 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 chemo-
type 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 work¹⁰ 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 dearoma-
tization 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

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

- 198 • $250 \leq MW \leq 750$;
- 199 • $\log P \leq 7$;
- 200 • number of H-bond acceptors ≤ 15 ;
- 201 • number of H-bond donors ≤ 8 ;
- 202 • number of rotatable bonds ≤ 15 .

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

206 **2.2. DNA-Encoded Libraries.** 1M representative subsets
207 for all 2497 DELs were generated in our previous work¹⁰ with
208 the help of the eDesigner tool.¹² This was done using
209 commercially available BBs from eMolecules and Enamine that
210 satisfy the Ro2¹³ and eDesigner built-in DNA-compatibility
211 filters. The enumerated compounds were standardized in the
212 same way as the ChEMBL dataset.

3. METHODS

213 **3.1. Generative Topographic Mapping.** In chemo-
214 informatics, each molecule can be represented as a data
215 point defined by a vector of numerical values called
216 descriptors. Molecules populate a chemical space, which is a
217 high-dimensional vector space. To analyze and comprehen-
218 sively visualize it, dimensionality reduction methods are
219 needed. GTM^{14–16} was the herein-used dimensionality
220 reduction tool. It works by fitting a manifold (flexible
221 hypersurface) into the multidimensional descriptor space
222 populated by “frame” items, followed by the projection of
223 the data points onto the thereupon defined 2D latent space
224 grid.

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

243 Different types of GTM *landscapes* can be created for the
244 same library, where properties of the compounds projected
245 onto each node are rendered using a color code. Three major
246 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 screen-
ing.^{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

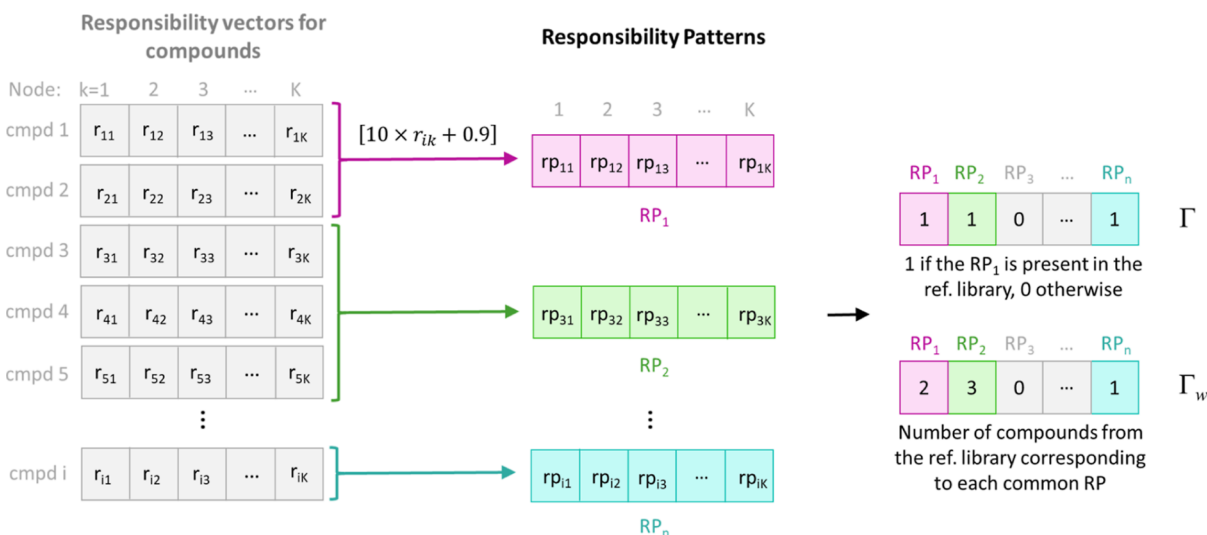


Figure 1. Summary of the RP-based library representations. The Γ values for a particular library are assigned based on the presence or absence of a certain RP in the reference library, and the Γ_w values represent the counts of reference library compounds covered by this RP.

309 compounds may not have exactly the same responsibility
 310 vector. However, similar compounds still are projected onto
 311 the map in a similar manner—according to a RP²⁰
 312 representing discretized responsibility vector according to eq 1

$$313 \quad rp_{ik} = [10 \times r_{ik} + 0.9] \quad (1)$$

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

317 It follows from eq 1 that responsibility values smaller than
 318 0.01 are reassigned to zero, and all others—to integer numbers
 319 from 1 to 10. Molecules situated close to each other in N -
 320 dimensional descriptor space and having slightly different
 321 responsibility vectors may have the same RP. These
 322 compounds usually share the same scaffold or substantial
 323 (connected or disconnected) maximum common substructure,
 324 or pharmacophore.²¹ Thus, in a way, an RP could be associated
 325 with a prevalent “chemotype”.

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

334 **3.3.2. Normalized CRVs (Φ).** A CRV = (c_1, c_2, \dots, c_k) is the
 335 vector encoding a library by the sum of responsibility values
 336 over all molecules of the library in each node of the map, as
 337 shown in eq 2. In other words, to some degree, this vector
 338 allows the encoding of a library by the number of compounds
 339 associated with each node of the corresponding GTM plot.
 340 Thus, the CRV mathematically describes compound distribu-
 341 tion over the 2D map and consequently over the chemical
 342 space of the library that this map visualizes. Considering that
 343 each area of the map is populated by a particular prevailing
 344 chemotype, the CRV is a crude indirect way of assessing the
 345 occurrences of different chemotypes in the library without
 346 actually defining them.

$$c_k = \sum_i^N r_{ik} \quad (2) \quad 347$$

where r_{ik} is responsibility value of the molecule i in the node k 348

The CRV is intrinsically dependent on the size of the library 349
 it encodes. Therefore, when collections of different sizes are 350
 compared in a context in which size differences are not 351
 relevant, c_k must be normalized by library size N according to 352
 eq 3. The resulting normalized CRV (Φ) encodes relative 353
 compound distribution over the chemical space of the analyzed 354
 collection. 355

$$\Phi_k = \frac{c_k}{N} \quad (3) \quad 356$$

3.3.3. Library-Modulated CRV (Λ). So far, the CRV and Φ 357
 consider all the chemical space zones (nodes) to be equally 358
 important in describing the library. However, some nodes may 359
 be more important—for example, the ones found to be highly 360
 populated by reference library compounds. For this purpose, 361
 the CRV of the analyzed library (a) can be modulated with 362
 respect to the compound distribution of another reference 363
 collection (r). The resulting library-modulated CRV (Λ) can 364
 be computed from the Φ of both collections, by calculating the 365
 fraction of compounds of the analyzed library in the total 366
 population of each node, as shown in eq 4. In Λ , a value $\Lambda_k = 0$ 367
 is assigned to all empty nodes in both analyzed and reference 368
 libraries, whereas for all non-empty nodes $1 \leq \Lambda_k \leq 2$ vary as a 369
 function of the fraction of compounds from the analyzed 370
 library in a given node. Nodes populated exclusively by 371
 compounds from r and a have value $\Lambda_k = 1$ and 2, 372
 respectively, whereas mixed nodes containing compounds 373
 from both libraries have values in the range $1 < \Lambda_k < 2$. 374

$$\Lambda_k = 1 + \frac{\Phi_k(a)}{\Phi_k(a) + \Phi_k(r)} \quad (4) \quad 375$$

where Λ_k is Λ value in a given non-empty node k for analyzed 376
 library a , whereas $\Phi_k(a)$ and $\Phi_k(r)$ are normalized cumulated 377
 responsibilities in the node k for the analyzed and reference 378
 library, respectively. 379

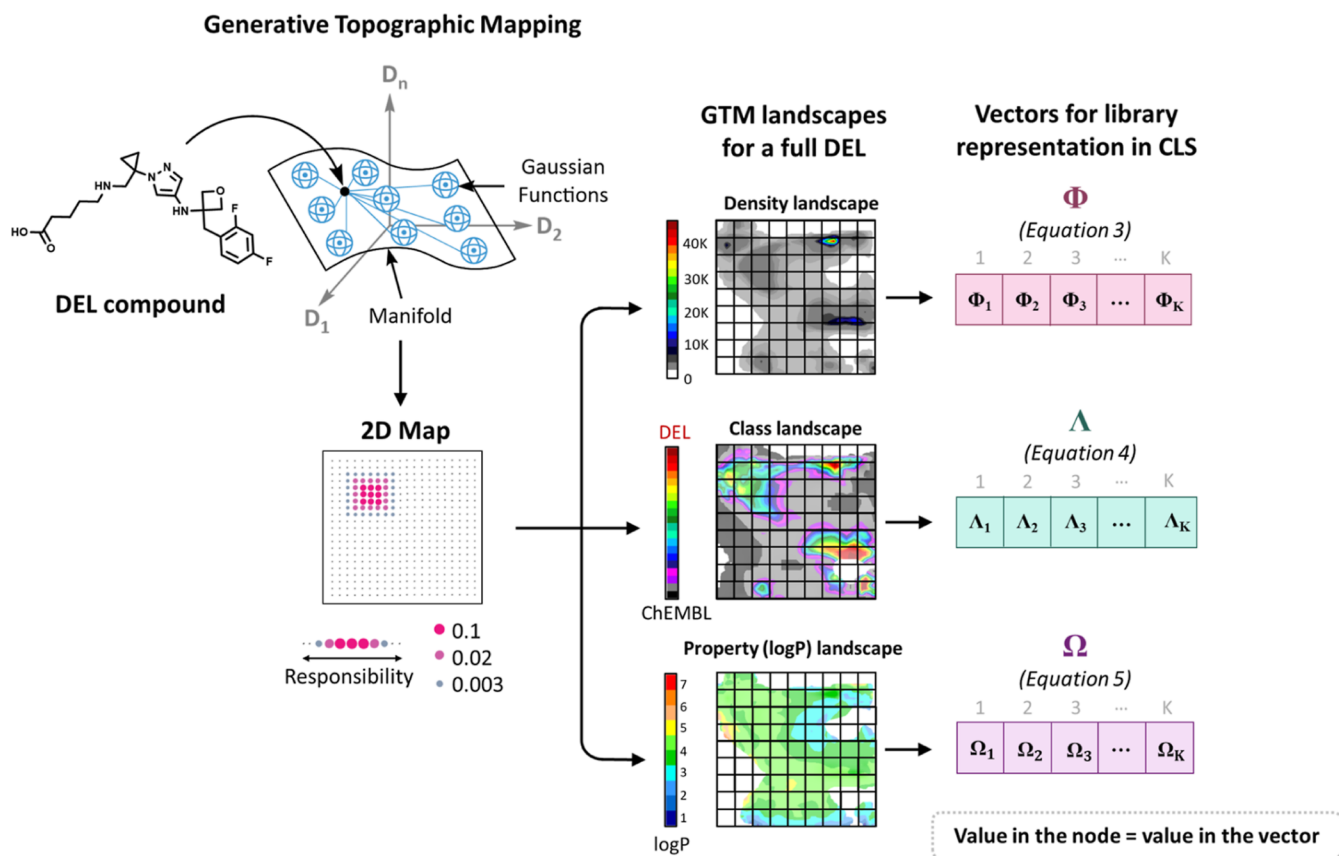


Figure 2. Scheme depicting how each of the introduced herein library encodings (Φ , Λ , and Ω) are derived from the GTM for a particular compound library.

380 When aiming to maximize representation and coverage of
 381 the reference collection by the analyzed library, the ideal case
 382 would be an Λ with $\Lambda_k = 0$ for the fully empty nodes and
 383 $\Lambda_k = 1.5$ (corresponding to equal representation of both
 384 reference and analyzed libraries) in all occupied ones. This
 385 “ideal” vector can thus be used as a reference in Tanimoto
 386 calculations for ranking libraries based on Λ .

387 **3.3.4. Property-Modulated CRV (Ω).** If the analysis of CLS
 388 should be performed in the context of some property or
 389 biological activity of underlying compounds for each library,
 390 the property-modulated CRV (Ω) can be used. Ω is composed
 391 of the mean property values for each node calculated according
 392 to eq 5.

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

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

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

405 **3.4. Similarity Relationships between Libraries in the**
 406 **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
 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.

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

where $\Gamma_{w_i}(r)$ 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 .

430 Notice that both coverage and weighted coverage scores
431 were used in our previous work¹⁰ for the comparison of virtual
432 DEL collections with the ChEMBL database.

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

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

437 Here, v is a chosen CRV-based representation ($v = \Phi, \Lambda, \Omega$),
438 and K is the total number of nodes.

4. RESULTS AND DISCUSSION

439 The herein proposed library encoding vectors Φ , Λ , Ω ,
440 Γ , and Γ_w provide different views of the CLS. To investigate
441 their usefulness, the pool of 2.5k previously generated DELs¹⁰
442 was used. Three case studies were performed. First, we
443 analyzed how proposed encodings and similarity metrics
444 handle the comparison of a large 88 M DEL with its 1 M
445 representative subset. The second case study addresses the
446 selection of the “optimal” DEL for the primary screening when
447 no or little information about the biological target is known.
448 The goal was to identify a DEL that covers “biologically
449 relevant” space (represented by ChEMBL) to the highest
450 extent. For this purpose, 2.5k DELs were compared to
451 ChEMBL (as a reference collection) in the CLS defined by
452 Γ , Γ_w , Φ , and Λ . In the third case study, the property-focused
453 analysis of the libraries was performed using the Ω encodings.

454 **4.1. Representative DEL Subset vs Its Parent Library:
455 A Test Study of Expected Near-Perfect Overlap.** In our
456 previous study,¹⁰ representative sets of each of the 2.5k DELs
457 were generated using random sampling of BBs in the
458 eDesigner¹² tool and not the full libraries. Such a sub-library
459 should be very similar to the entire DEL and cover virtually all
460 of its chemical space. Therefore, overlap analysis of a
461 representative DEL subset with respect to its parent library is
462 a baseline case for illustrating how well each of the encodings
463 reflects its close relationship.

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

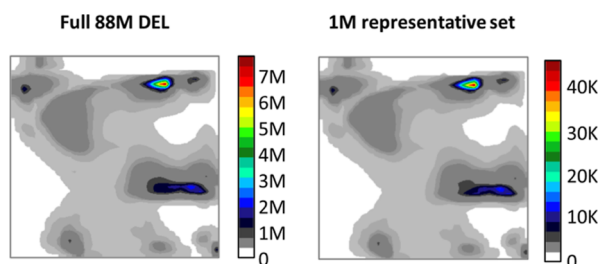


Figure 3. Density landscapes of the entire 88M DEL2568 and its 1M representative subset.

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
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—combina-
tions 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 $\Lambda_k = 1.5$, 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

Pairwise coverage (or similarity) of the reference (ChEMBL) chemical space by the analyzed libraries (2.5K DELs)

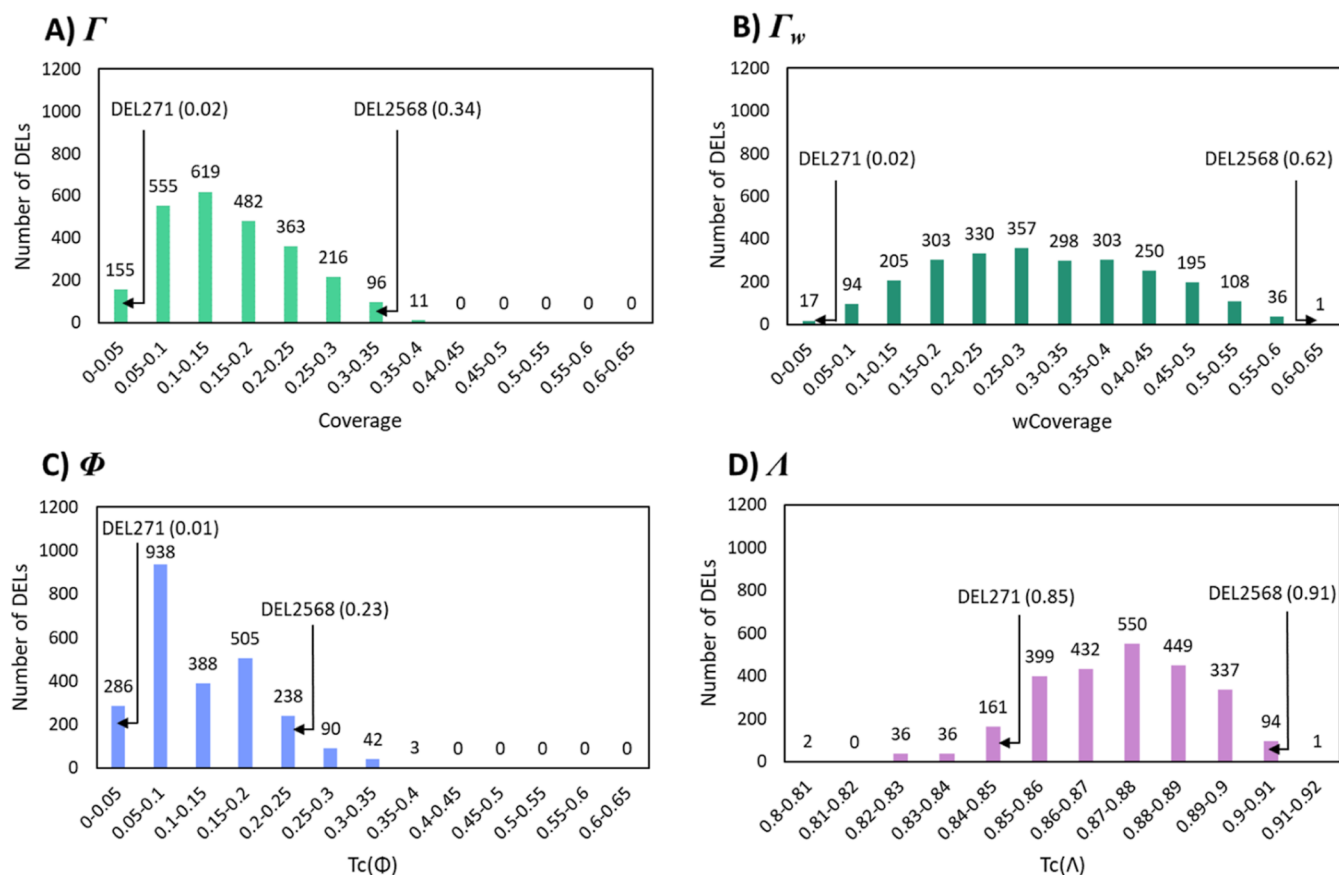


Figure 4. Pairwise comparison of 2.5k DELs with ChEMBL using different representations and metrics: distribution of ChEMBL coverage scores calculated using Γ (A) and Γ_w (B), and distribution of Tc between ChEMBL and each DEL calculated using Φ (C) and Λ (D).

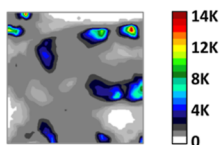
524 based on Γ_w —were selected as points of reference, to trace
 525 their scoring with other representations. Both Γ -based
 526 coverage (Figure 4A) and Φ -based Tc (Figure 4C) adopt
 527 values within a similar and rather low value range: from 0.01 to
 528 0.4. This highlights that DEL compound distribution is quite
 529 different from that of ChEMBL, and the likelihood of finding
 530 the ChEMBL RPs in DELs is rather low. However, the Γ_w -
 531 based coverage shows that those RPs that are covered by DELs
 532 in fact correspond to the prevailing compound population of
 533 ChEMBL because observed values of coverage almost doubled
 534 with respect to Γ -based coverage (Figure 4B). In all three
 535 cases, the two “marker” libraries, nevertheless, keep their
 536 relative rank: DEL2568 is always ranked in the top 5–10% of
 537 libraries and DEL271—in the last 10–15%. As expected,
 538 tuning the overlap criterion by means of the usage of different
 539 CLS vectors should never override the fundamental “core”
 540 library similarity, distinguishing between libraries containing
 541 closely related molecules from those which do not.

542 In the case of Λ -based similarity, the Tc values are spread
 543 within a narrow range: from 0.8 to 0.92 (Figure 4D). The Λ -
 544 based similarity spectrum is intrinsically different from those
 545 calculated using other encodings. Since vectors for all libraries
 546 are modulated with the CRV of the same reference collection,
 547 the similarity value between two Λ is always higher than that in
 548 the case of Φ , for example. However, the position of DEL2568
 549 and DEL271 in Figure 4D is similar to the other three cases.

Thus, even though being shifted toward higher values, 550
 similarity distribution in the CLS defined by Λ follows the 551
 same trends as in other library spaces. 552

For further analysis of the similarity relationships in the four 553
 proposed representations of CLS, all DELs were ranked with 554
 respect to the coverage of (or similarity to) ChEMBL. To 555
 simplify the analysis, here we analyze only five DELs: ranked 556
 the first, 50th, 100th, 1000th, and 2497th with respect to 557
 ChEMBL. For each of these five DELs, a density landscape 558
 showing compound distribution in the chemical space of the 559
 library was created (see Figure 5). This figure shows that each 560
 of the representations ranks libraries differently—none of the 561
 libraries were selected as the best one by more than one 562
 representation. However, DELs having the same rank in 563
 different spaces (landscapes forming columns in Figure 5) still 564
 have very similar compound distribution over the map. Failure 565
 to consensually score one DEL as the best match for ChEMBL, 566
 in any CLS, is due to the fact that there are several DELs that 567
 might claim this title, and no single one is undoubtedly 568
 outstanding in terms of sharing related chemotypes with 569
 ChEMBL. Looking at the problem through the prism of 570
 multiple CLS definitions is evidencing this important aspect, 571
 that is, allowing for more flexibility in experimental setups. In 572
 this scenario, there is no particular reason to pick either of the 573
 DELs of column no 1 in Figure 5—a case in which extraneous 574
 parameters (availability, facility of synthesis, and cost) may be 575

ChEMBL Density landscape



Density landscapes of DELs ranked by their coverage and similarity to ChEMBL

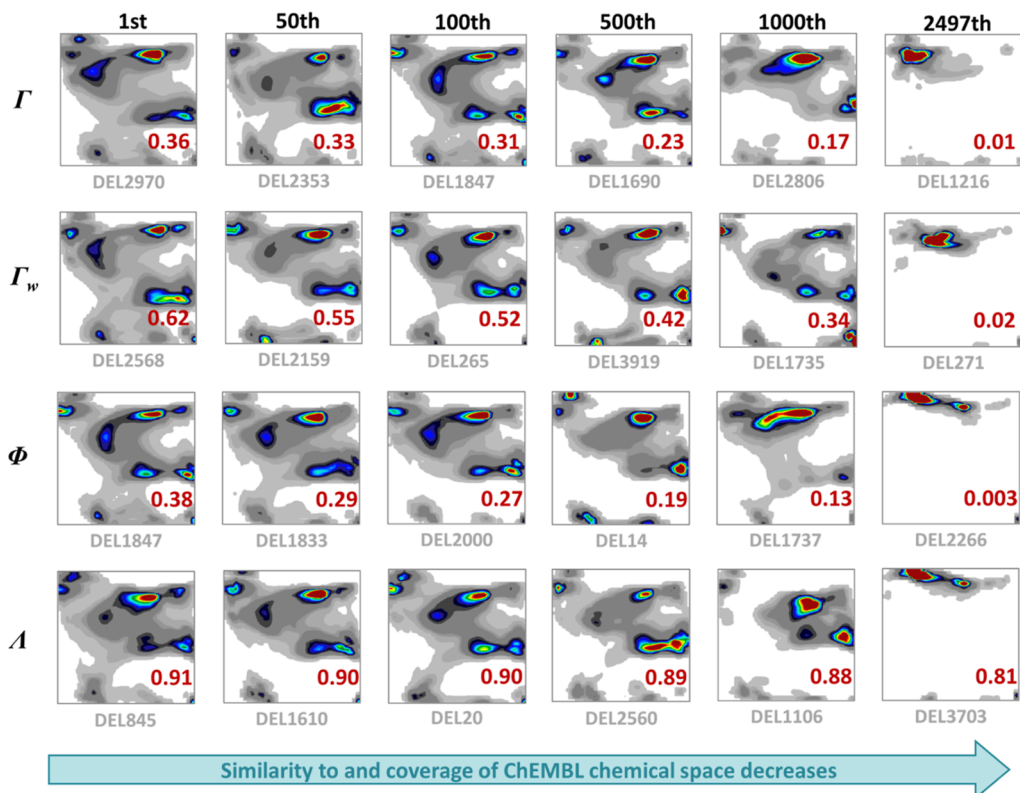


Figure 5. Density GTM landscapes of ChEMBL28 and selected DELs ranging from the most similar to the least similar to ChEMBL. DELs were selected and ranked either by coverage scores (in the case of Γ and Γ_w) or Tanimoto similarity coefficients (in the case of Φ and Λ). Values of either coverage or Tc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

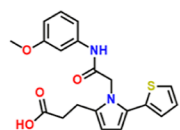
576 applied by the user to select either of these. Should a
 577 consensual winner emerge from this analysis, selecting it at
 578 higher costs over the others may make sense. Practically,
 579 however, visual inspection shows that the first few hundred
 580 DELs have similar density landscapes to the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

590 Even though each of the analyzed representations offers a
 591 different DEL as the closest to ChEMBL (DEL2970,
 592 DEL2568, DEL1847, and DEL845), they all appear to be
 593 quite similar. Interestingly, all these libraries are three-cycled
 594 DELs that were designed exclusively based on robust coupling
 595 reactions—aldehyde reductive amination (all four libraries),
 596 Ullmann-type *N*-aryl coupling (DEL2970 and DEL845),

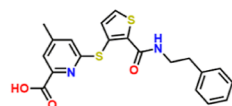
Migita thioether synthesis from thiophenols and arylbromides
 (DEL1847 and DEL2568), and carboxylic acid/amine
 condensation (DEL1847 and DEL845) (see Figure S2 of
[Supporting Information](#)). The size of the full DELs is also very
 similar for those four libraries—slightly above 80M com-
 pounds. The 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, 617

A1. Thiophene-containing compounds

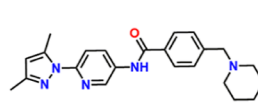


CHEMBL1386410

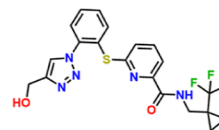


DEL1847-8440

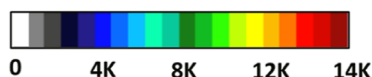
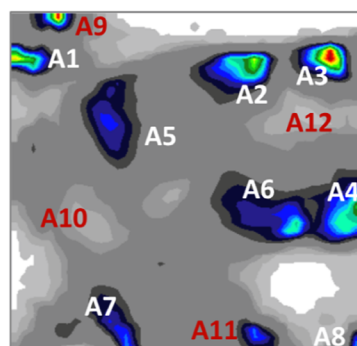
A2. Azoles-containing compounds



CHEMBL4549155



DEL1847-829238

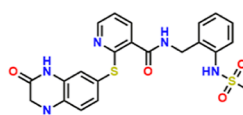


Filtered ChEMBL
(reference collection)

A3. Sulfonanilides

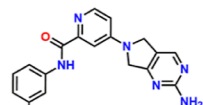


CHEMBL1651007

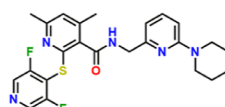


DEL1847_453242

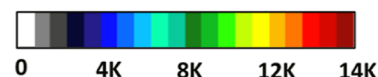
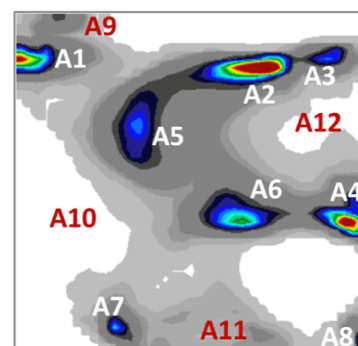
A4. Halogenated N-heterocyclic compounds



CHEMBL3702639

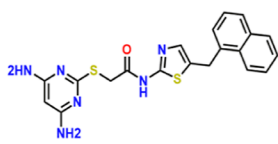


DEL1847-8678

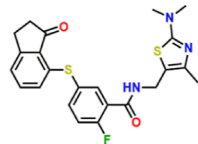


DEL1847
($Tc(\Phi)=0.38$)

A9. 2-Aminothiazole-containing compounds

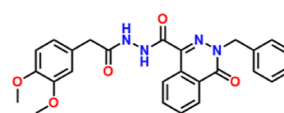


CHEMBL3771201

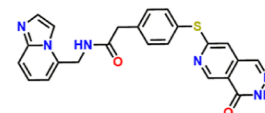


DEL1847_45654

A11. Pyridazinone/oxadiazolone containing amides



CHEMBL1385968



DEL_1847-22798

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.

618 which means that their diversity is much lower, and those
619 DELs can be considered as focused libraries containing very
620 similar compounds. This is explainable by the fact that
621 employing two heterocyclization steps in DEL synthesis means
622 that all compounds possess at least two identical hetero-
623 cycles—a consequently large scaffold—with diversity being
624 introduced only *via* their “ornaments”, by contrast to, say, an
625 amide formation in which everything but the $-C(=O)NH-$
626 moiety is variable.

627 The use of only heterocyclizations is convenient for
628 “focused” DEL synthesis, as the common scaffold generated
629 by the reaction represents a common signature of all library
630 members, which vary in terms of scaffold substituents only.²³
631 This provides an excellent library for extracting structure–
632 activity relations and fine-tuning lead molecules, provided, of
633 course, that the focus around the chosen heterocyclic core
634 matches the actual chemical space zone favored by the target.
635 However, if the goal is to produce general-purpose DELs, it is a
636 safer option to use building-block-rich coupling reactions
637 instead because abundant BB classes exist. Many BBs already
638 contain necessary heterocyclic moieties,²⁴ albeit not necessarily
639 connected to each other in a same way as they would be linked

up in a heterocyclization synthesis-based DEL. Another option
640 might be to use only one heterocyclization step combined with
641 two coupling synthetic cycles. In this way, the diversity coming
642 from coupling reactions can partially compensate for the
643 presence of the same heterocycle in each molecule. An example
644 of such design is DEL2806 (1000th library by Γ)—it combines
645 imidazole synthesis with guanidine group formation from
646 amines and Ullmann-type *N*-aryl coupling. All other DELs
647 featuring from 1st to 500th in Figure 5 are based only on
648 coupling reactions.

4.2.2. *In-Depth Analysis and Interpretability of Library*
649 *Overlap.* Overlap scores are useful for the rapid processing and
650 ranking of large sets of candidate libraries, but a real
651 understanding of overlap must go down to individual
652 compound structure levels. The strength of this protocol is
653 that the mapping used to define CLS vectors can implicitly
654 support this approach. To illustrate that, the density landscape
655 for DEL1847 that is the closest to ChEMBL according to Φ
656 ranking was compared to the density landscape of ChEMBL
657 (Figure 6). DEL1847 is a three-step library based on aldehyde
658 reductive amination with the NH_2 group of the headpiece
659 (2652 aldehydes), followed by the condensation of the same
660

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

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

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

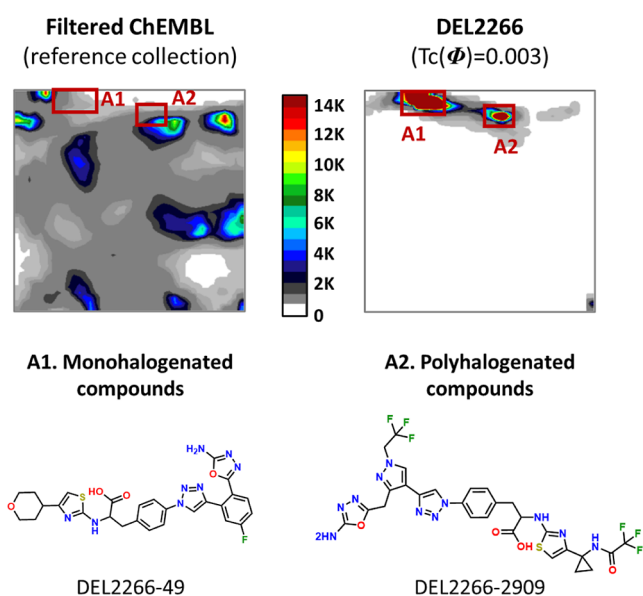


Figure 7. Interpretation of the similarity between ChEMBL and DEL2266 via structural analysis of the density landscapes of these libraries.

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 conven-
 tional 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 $T_c = 0.95$. 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 $T_c = 0.78$.
 The T_c 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 ($T_c(\Phi) = 0.13$).
 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 ($T_c(\Phi) = 0.34$). Furthermore, the
 local property distribution in this collection is much closer to
 ChEMBL than that in DEL2189. Indeed, there are many areas

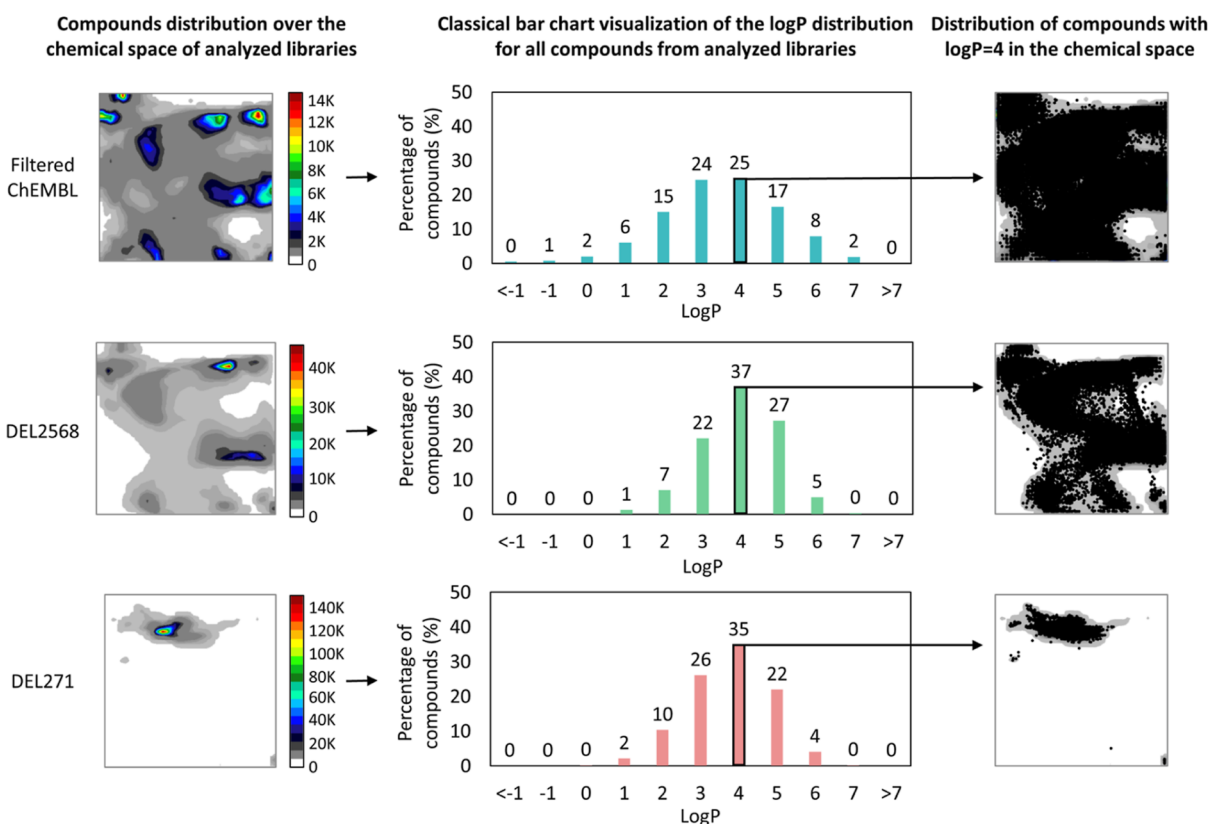


Figure 8. (Left) Density landscapes of filtered ChEMBL, DEL2568, and DEL271; (center) classical bar chart visualization of calculated log P distribution for all compounds from analyzed libraries; (right) compounds with log $P = 4$ (black dots) projected on the corresponding density landscapes.

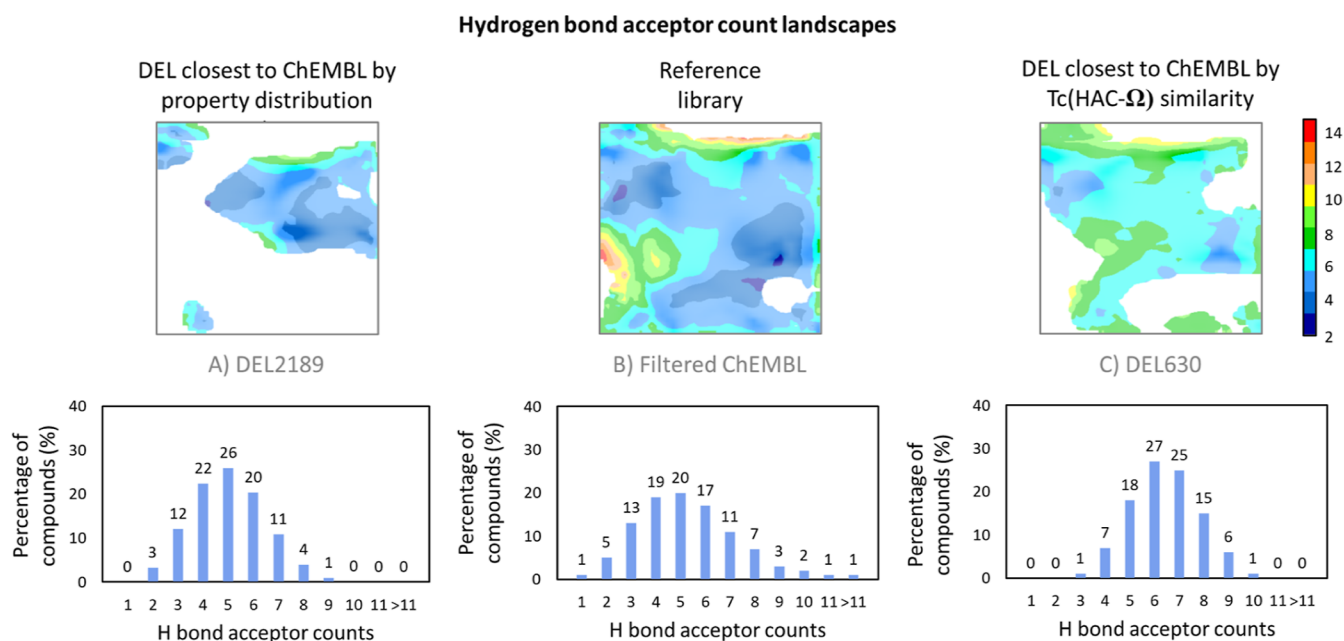


Figure 9. Hydrogen bond acceptor count (HAC) landscapes for (A) DEL2189 (selected by property distribution similarity), (B) reference library-filtered ChEMBL28, and (C) DEL630 [selected by Tc(HAC- Ω) similarity].

763 colored in the same way in both ChEMBL and DEL630
 764 collections, which means that the average number of H-bond
 765 acceptors in compounds populating these zones is very close.
 766 Thus, Ω encoding allows us to take into consideration both
 767 property and chemotype distribution in the chemical space of

analyzed libraries. Different Ω can be created using any 768
 measured or calculated property if it is provided for every 769
 compound in analyzed libraries. Figure S4 renders the 770
 distribution of the similarity of DELs with respect to ChEMBL 771
 in six Ω -encoded CLS: MW, log P , H-bond acceptors and 772

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

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

5. CONCLUSIONS

In this work, we reported the development of several types of 799
 vector-based encodings for characterizing libraries of various 800
 sizes and compositions as a function of the relative distribution 801
 of molecules in the GTM-based chemical space. These 802
 representations constitute a new way of the analysis of 803
 combinatorial mixtures, such as DELs, that should be 804
 considered not only as an ensemble of compounds, but also 805
 as unified entities—mixtures whose composition cannot be 806
 easily changed once synthesized. Of course, the methodology 807
 generally applies in contexts where any library—cherry- 808
 pickable or not—needs to be regarded as a stand-alone entity, 809
 rather than a collection of individual molecules. With the 810
 encodings introduced here, it becomes possible to clearly 811
 define CLS where each collection is considered as a data point. 812
 Classical cheminformatics allows for the management of a 813
 portfolio of *compounds* forming a core library (comparison to 814
 other compound sets, directed enrichment in new compounds, 815
 focused subset extraction for screening, *etc.*), whereas this 816
 methodology enables the management of a portfolio of *libraries* 817
 (selection of the best suited one for a screening campaign, 818
 enrichment with novel libraries—overlapping or not, *etc.*). 819

From the example of ChEMBL *vs* DEL comparison, it was 820
 shown that all proposed CLS representations—responsibility 821
 pattern fingerprints (Γ), responsibility count vectors (Γ_w), 822
 normalized CRVs (Φ), library-modulated CRVs (Λ), and 823
 property-modulated CRVs (Ω)—are able to efficiently encode 824
 key information about the “chemotype” distribution of 825
 analyzed libraries, where “chemotypes” are implicitly defined 826
 by the intrinsic neighborhood compliance of GTMs. “chemo- 827
 types”, in this sense, may be common scaffolds including or not 828
 common key “ornaments”, common topological pharmaco- 829
 phores, or more loosely defined compound clusters of 830
 molecules with a specific global charge or outstanding size, 831
etc. Similarity relationships in all five CLSs seem reasonable 832
 and chemically meaningful and allow adequate sorting of DELs 833

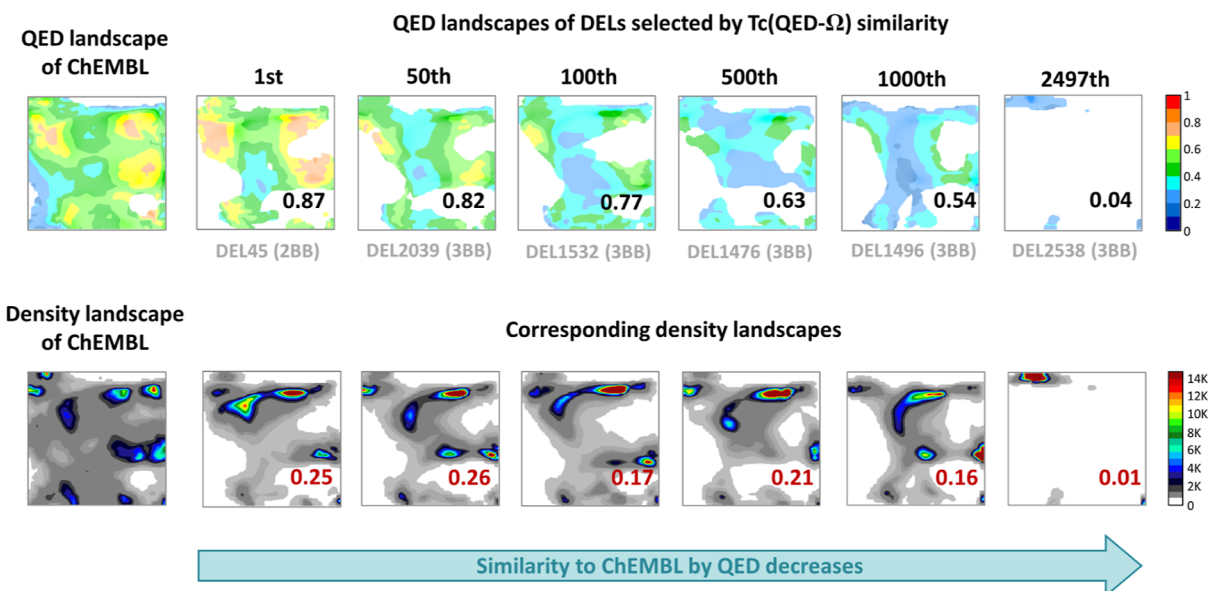


Figure 10. First row: On the left: QED landscape of filtered ChEMBL28. On the right: QED landscapes of DELs ranging from the most similar to the least similar to ChEMBL sorted by their Tanimoto coefficients calculated based on their QED- Ω with respect to ChEMBL (in black). Second row: On the left: density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

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

849 To analyze libraries with respect to the featured chemotypes
850 without paying attention to their population the best choice
851 would be Γ . If the population of the matched chemotypes in
852 only one of the libraries (reference collection) is important—
853 the coverage score based on the Γ_w should be used, thereby
854 ensuring that the candidate library matches the often-seen
855 patterns in the reference collection, and not its atypical
856 “singletons”. In case the compound distribution over the
857 chemical space of all analyzed collections is important, CLS
858 should be defined by the Φ , whereas Tanimoto similarity
859 should be used for library ranking. This strategy can also be
860 used in order to select a library that maximally reproduces
861 compound distribution from the chemical space of the
862 reference collection (e.g., selection of the optimal representa-
863 tive subset). Λ -based encoding is particularly useful when one
864 wants to compare a coverage of a reference dataset by some
865 other libraries. In this case, each library is encoded considering
866 its relative compounds distribution with respect to the
867 reference collection, so a special accent is placed on the
868 differences between the relative proportion of compounds
869 coming from analyzed and reference libraries without taking
870 into consideration the absolute popularity of each node.
871 Moreover, in case the accent of the analysis is placed on the
872 particular calculated or measured property, Ω can be used to
873 encode libraries with respect to both chemotype and property
874 distribution in the chemical spaces of these collections. In
875 contrast to classical property histograms that describe the
876 global distribution of the property values among compounds of
877 the whole library, Ω encodes local property distribution among
878 compounds belonging to different chemotypes and populating
879 particular areas of the chemical space.

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

891 Now, when the performance of the proposed encodings and
892 the similarity behavior of libraries (objects) in corresponding
893 CLS are analyzed and described, it should be last but not least
894 noted that this CLS may also be visualized, like any “classical”
895 chemical space. In perspective, the meta-GTM approach²⁹ is

perfectly suited for the dimensionality reduction and visual- 896
ization of CLS. 897

■ ASSOCIATED CONTENT 898

SI Supporting Information 899

The Supporting Information is available free of charge at 900
<https://pubs.acs.org/doi/10.1021/acs.jcim.3c00520>. 901

Venn diagrams comparing several DELs, density land- 902
scapes of selected DELs, and distributions of some 903
physicochemical parameters of selected DELs (PDF) 904

■ AUTHOR INFORMATION 905

Corresponding Author 906

Alexandre Varnek — *Laboratory of Chemoinformatics,* 907
University of Strasbourg, Strasbourg 67081, France; 908
orcid.org/0000-0003-1886-925X; Phone: +33 909
368851560; Email: varnek@unistra.fr 910

Authors 911

Regina Pikalyova — *Laboratory of Chemoinformatics,* 912
University of Strasbourg, Strasbourg 67081, France 913
Yuliana Zabolotna — *Laboratory of Chemoinformatics,* 914
University of Strasbourg, Strasbourg 67081, France 915
Dragos Horvath — *Laboratory of Chemoinformatics,* 916
University of Strasbourg, Strasbourg 67081, France; 917
orcid.org/0000-0003-0173-5714 918
Gilles Marcou — *Laboratory of Chemoinformatics, University* 919
of Strasbourg, Strasbourg 67081, France; [orcid.org/](https://orcid.org/0000-0003-1676-6708) 920
[0000-0003-1676-6708](https://orcid.org/0000-0003-1676-6708) 921

Complete contact information is available at: 922
<https://pubs.acs.org/10.1021/acs.jcim.3c00520> 923

Funding 924

R.P.—Bourse de l’Ecole Doctorale des Sciences Chimiques 925
ED222, Université de Strasbourg. 926

Notes 927

The authors declare no competing financial interest. 928
The data used in this work are available in the public domain 929
resources: biologically relevant compounds from ChEMBL³⁰ 930
(version 28)—<https://www.ebi.ac.uk/chembl/>, eMolecules³¹ 931
BBs that were used for DEL generation using eDesigner¹² are 932
partially available on the website [https://www.emolecules.](https://www.emolecules.com/products/building-blocks) 933
[com/products/building-blocks](https://www.emolecules.com/products/building-blocks), and Enamine³² BBs are avail- 934
able on the website <https://enamine.net/building-blocks>. 935

■ ACKNOWLEDGMENTS 936

The authors would like to thank eMolecules Inc.³¹ for 937
providing the collection of commercially available BBs that 938
were used for the generation of DELs analyzed in this work. 939

■ REFERENCES 940

- (1) Czarnik, A. W. Encoding methods for combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1997**, *1*, 60–66. 941
- (2) Brenner, S.; Lerner, R. A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5381–5383. 942
- (3) Franzini, R. M.; Neri, D.; Scheuermann, J. DNA-encoded 943
chemical libraries: advancing beyond conventional small-molecule 944
libraries. *Acc. Chem. Res.* **2014**, *47*, 1247–1255. 945
- (4) Fourches, D.; Tropsha, A. Using graph indices for the analysis 946
and comparison of chemical datasets. *Mol. Inf.* **2013**, *32*, 827–842. 947
- (5) Miranda-Quintana, R. A.; Bajusz, D.; RÁCZ, A.; Héberger, K. 948
Extended similarity indices: the benefits of comparing more than two 949
950
951

- 952 objects simultaneously. Part 1: Theory and characteristics. *J. Cheminf.*
953 **2021**, *13*, 32.
- 954 (6) Dunn, T. B.; Seabra, G. M.; Kim, T. D.; Juárez-Mercado, K. E.;
955 Li, C.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Diversity and
956 Chemical Library Networks of Large Data Sets. *J. Chem. Inf. Model.*
957 **2021**, *62*, 2186–2201.
- 958 (7) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.;
959 Medina-Franco, J. L. Consensus diversity plots: a global diversity
960 analysis of chemical libraries. *J. Cheminf.* **2016**, *8*, 63.
- 961 (8) Fernández-de Gortari, E.; García-Jacas, C. R.; Martínez-
962 Mayorga, K.; Medina-Franco, J. L. Database fingerprint (DFP): an
963 approach to represent molecular databases. *J. Cheminf.* **2017**, *9*, 9.
- 964 (9) Bishop, C. M.; Svensen, M.; Williams, C. K. I. GTM: The
965 generative topographic mapping. *Neural Comput.* **1998**, *10*, 215–234.
- 966 (10) Pikalyova, R.; Zabolotna, Y.; Volochnyuk, D. M.; Horvath, D.;
967 Marcou, G.; Varnek, A. Exploration of the Chemical Space of DNA-
968 encoded Libraries. *Mol. Inf.* **2022**, *41*, 2100289.
- 969 (11) ChemaAxon. *JChem*, version 20.8.3; ChemAxon Ltd: Budapest,
970 Hungary, 2020.
- 971 (12) Martin, A.; Nicolaou, C. A.; Toledo, M. A. Navigating the DNA
972 encoded libraries chemical space. *Commun. Chem.* **2020**, *3*, 127–129.
- 973 (13) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.;
974 Tomkinson, N. P. Designing novel building blocks is an overlooked
975 strategy to improve compound quality. *Drug Discovery Today* **2015**,
976 *20*, 11–17.
- 977 (14) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D.
978 Mappability of drug-like space: towards a polypharmacologically
979 competent map of drug-relevant compounds. *J. Comput.-Aided Mol.*
980 *Des.* **2015**, *29*, 1087–1108.
- 981 (15) Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk,
982 D. M.; Varnek, A. Chemography: Searching for Hidden Treasures. *J.*
983 *Chem. Inf. Model.* **2021**, *61*, 179–188.
- 984 (16) Lin, A. Cartographie topographique générative: un outil
985 puissant pour la visualisation, l'analyse et la modélisation de données
986 chimiques volumineuses. Ph.D. Thesis, Université de Strasbourg:
987 Strasbourg, 2019.
- 988 (17) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath,
989 J.; Varnek, A. Virtual screening with generative topographic maps:
990 how many maps are required? *J. Chem. Inf. Model.* **2018**, *59*, 564–572.
- 991 (18) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA
992 Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855–
993 868.
- 994 (19) Horvath, D.; Marcou, G.; Varnek, A. Generative topographic
995 mapping in drug design. *Drug Discovery Today: Technol.* **2019**, *32–33*,
996 99–107.
- 997 (20) Klimenko, K.; Marcou, G.; Horvath, D.; Varnek, A. Chemical
998 Space Mapping and Structure-Activity Analysis of the ChEMBL
999 Antiviral Compound Set. *J. Chem. Inf. Model.* **2016**, *56*, 1438–1454.
- 1000 (21) Kayastha, S.; Kunimoto, R.; Horvath, D.; Varnek, A.; Bajorath,
1001 J. From bird's eye views to molecular communities: two-layered
1002 visualization of structure–activity relationships in large compound
1003 data sets. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 961–977.
- 1004 (22) Zabolotna, Y.; Volochnyuk, D. M.; Ryabukhin, S. V.; Horvath,
1005 D.; Gavrilenko, K. S.; Marcou, G.; Moroz, Y. S.; Oksiuta, O.; Varnek,
1006 A. A close-up look at the chemical space of commercially available
1007 building blocks for medicinal chemistry. *J. Chem. Inf. Model.* **2021**, *62*,
1008 2171–2185.
- 1009 (23) Dickson, P.; Kodadek, T. Chemical composition of DNA-
1010 encoded libraries, past present and future. *Org. Biomol. Chem.* **2019**,
1011 *17*, 4676–4688.
- 1012 (24) Oksiuta, O. V.; Pashenko, A. E.; Smalii, R. V.; Volochnyuk, D.
1013 M.; Ryabukhin, S. V. Heterocyclization vs Coupling Reactions: A
1014 DNA-Encoded Libraries Case. *Zh. Org. Farm. Khim.* **2023**, *21*, 3–19.
- 1015 (25) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.;
1016 Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P. Drug-like
1017 annotation and duplicate analysis of a 23-supplier chemical database
1018 totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*,
1019 643–651.
- (26) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, 1020
A. Drug-and lead-likeness, target class, and molecular diversity 1021
analysis of 7.9 million commercially available organic compounds 1022
provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470–479. 1023
- (27) Lucas, X.; Gruning, B. A.; Bleher, S.; Gunther, S. The 1024
purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* 1025
2015, *55*, 915–924. 1026
- (28) Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K.-C. 1027
Assessment of chemical libraries for their druggability. *Comput. Biol.* 1028
Chem. **2005**, *29*, 55–67. 1029
- (29) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. 1030
Chemical data visualization and analysis with incremental generative 1031
topographic mapping: big data challenge. *J. Chem. Inf. Model.* **2015**, 1032
55, 84–94. 1033
- (30) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, 1034
M.; Félix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; 1035
Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; 1036
Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; 1037
Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: 1038
towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, 1039
D930–D940. 1040
- (31) eMolecules Inc. <https://www.emolecules.com/>. 1041
- (32) Enamine Ltd. <https://enamine.net/>. 1042