UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

Chimie de la matière complexe – UMR 7140

THÈSE présentée par :

Regina PIKALYOVA

soutenue le : **12 septembre 2024**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie informatique et théorique

Chémoinformatique des chimiothèques à codage ADN : design, génération in silico, gestion, analyse, et comparaison

THÈSE dirigée par :

M. VARNEK Alexandre	Professeur, Université de Strasbourg
M. HORVATH Dragos	Directeur de recherche CNRS, Université de Strasbourg

RAPPORTEURS :

Mme. DOUGUET Dominique	Chargée de recherche Inserm (HDR), Université Côte d'Azur
M. MORELLI Xavier	Directeur de recherche, CNRS, Université d'Aix-Marseille

Acknowledgments

I would like to express my deepest gratitude to my thesis supervisors, Professor Alexandre Varnek and Dr. Dragos Horvath, for their exceptional guidance and support. Working under your expert supervision has been an incredibly enriching experience, and I am profoundly grateful for the knowledge and insights you have shared with me.

I am also immensely thankful to the esteemed members of the jury, Dr. Dominique Douguet and Dr. Xavier Morelli, for dedicating their valuable time to read and evaluate my work.

My heartfelt thanks go to my colleagues and friends in the laboratory, especially Pierre, Alexey, Farah, Tagir, Maxim, Helena, Sai, Polina, and Shamkhal, whose friendship and kindness have made this journey even more rewarding. A special mention goes to Tagir, with whom I had the privilege of working on our final project. Tagir, working and collaborating with you was an enriching and refreshing experience.

I would also like to thank all laboratory members whose everyday support was invaluable. I am deeply grateful to Dr. Olga Klimchuk, Dr. Fanny Bonachera, and Dr. Gilles Marcou for their friendliness, responsiveness, and willingness to help with any issues. I would also like to extend my thanks to my dear friends Lena, Elmira, Rashid, and colleagues Yuliana, Kyrylo, Iryna, Arkadii, Mikhail, Louis, William, Aikhan, and others.

Of course, my deepest appreciation goes to my family, particularly my twin sister and my parents, who have always believed in me and supported me unconditionally. Your unwavering faith in my abilities has been a constant source of strength and motivation.

I am profoundly grateful to my sweetheart, Vincent, for his constant emotional support and unconditional love. Your encouragement and patience have helped me stay focused and resilient during the most challenging times.

I would also like to thank all professors from the faculty of chemistry at the University of Strasbourg. Your exceptional pedagogy and daily inspiration have been instrumental in shaping my understanding and passion for chemistry.

Finally, I dedicate this thesis to Dr. Sholpan Urkendovna Aliyeva, whose daily efforts in saving women's lives are truly heroic. Your dedication and compassion have been an inspiration to me.

Table of Contents

1. Résumé en français 3

1.1 Introduction 3

1.2 Résultats et discussions 6

1.2.1 Exploration de l'espace chimique de DELs et sélection d'une chimiothèque optimale pour le criblage primaire 6

1.2.2 Espace vectoriel des chimiothèques 8

1.2.3 Meta-GTM : visualisation de l'ensemble des chimiothèques 10

1.2.4 Analyse de la DEL ciblée contre la protéine BRD4 12

1.2.5 Visualisation de l'espace d'une chimiothèque combinatoire sur GTM sans énumération explicite des structures à l'aide de l'apprentissage profond 13

1.3 Conclusion générale 15

1.4 Liste des présentations 17

1.5 Liste des publications 18

2. Introduction 19

2.1 Need for Alternative Compound Screening Technologies 19

2.2 DNA-Encoded Library (DEL) Technology 21

2.3 DEL chemoinformatics and its gaps 25

2.3.1 Building Block selection 25

2.3.2 Enumeration 26

2.3.3 DEL chemical space analysis and comparison 28

2.3.4 Property analysis 29

2.3.5 DEL hit analysis 29

2.3.6 Overview of the gaps in DEL chemoinformatics 30

2.4 Methods of comparison/analysis of chemical libraries 33

2.5 Generative Topographic Mapping (GTM) for chemical library analysis and comparison 39

2.5.1 Generative Topographic Mapping (GTM) 39

2.5.2 Universal GTM 40

2.5.3 Chemical Space (CS) and Chemical Library Space (CLS) 40

2.5.4 CLS vectors and GTM landscapes 41
2.5.5 Responsibility Patterns 44
2.5.6 Meta-GTM 47
2.6 ML modeling for DEL hit prioritization and BB reactivity prediction 51
2.6.1 SVM and SVR 51
3. Thesis outline 55
4. Exploration of the chemical space of DNA-Encoded Libraries 57
Introduction 57
Summary 75
5. Chemical Library Space (CLS) analysis 77
Introduction 77
Summary 93
6. Meta-GTM: a tool for Chemical Library Space visualization 95
Introduction 95
Summary 109
7. BB reactivity prediction and hit prioritization: BRD4 focused DEL study 111
Introduction 111
Summary 129
8. Combinatorial Library Network (CoLiNN) for combinatorial library visualization without compound enumeration 131
Introduction 131
Summary 153
9. General conclusion and perspectives 155
10. Abbreviations 159
11. References 161

1. Résumé en français

1.1 Introduction

La technologie des chimiothèques codées par l'ADN[1] (*DNA-Encoded Libraries,* DELs) est l'une des méthodes de criblage récentes utilisées dans la découverte de médicaments. Elle permet de découvrir des molécules organiques qui se lient de manière non covalente à une cible biologique particulière et donc présentent un effet biologique souhaité. La DEL est une collection combinatoire pouvant aller jusqu'à 10^{12} de composés[2]. Chaque composé dans cette chimiothèque est attaché de manière covalente à une étiquette d'ADN. Cette dernière encode des informations sur les éléments constitutifs à partir desquels la molécule a été synthétisée. D'une certaine manière, l'étiquette d'ADN joue le rôle d'un « code-barres » qui permet d'encoder des informations structurelles sur la molécule.

Le criblage des composés encodés par ADN est effectué par la sélection par affinité contre une cible biologique donnée pour identifier les composés prometteurs. Ce processus diffère considérablement du criblage à haut débit (*High Throughput Screening*, *HTS*[3]) conventionnel des chimiothèques. Les composés DEL sont mélangés avec la protéine cible immobilisée sur un support solide dans un seul récipient où tous les composés sont en compétition pour se lier à la cible biologique. Les molécules qui ne parviennent pas à se lier à la protéine cible sont éliminées tandis que les ligands à haute affinité sont séparés de la protéine. Les étiquettes d'ADN qui encodent les ligands à haute affinité sont ensuite amplifiées en utilisant la technologie PCR et décodées par séquençage ADN ce qui permet d'identifier les structures des composés prometteurs[4].

La technologie DEL présente de nombreux avantages, tant pour l'industrie que pour les laboratoires universitaires. Comme les DELs sont synthétisées en utilisant une approche combinatoire de division et de regroupement (*Split and Pool*), cette technologie permet de produire plusieurs chimiothèques de taille énorme[4]. Les composés DEL sont criblés tous à la fois dans un seul récipient permettant d'explorer en une fois des larges régions de l'espace chimique[5]. La configuration expérimentale simple de la sélection par

affinité, accessible tant dans les laboratoires industriels qu'universitaires permet une identification rapide et peu coûteuse des composés prometteurs[4]. De nombreuses histoires de succès de l'utilisation de cette technologie ont été publiées, y compris des composés dérivés des DEL qui ont progressé jusqu'aux essais cliniques[4]. La technologie de la DEL représente donc un outil précieux pour les projets de découverte de médicaments.

Jusqu'à présent, la plupart des efforts de recherche computationnelle dans le domaine de la DEL se sont concentrés sur l'analyse des résultats du criblage par affinité[6,7], à savoir sur l'analyse des ligands de la protéine en question, alors que la préparation initiale de la DEL est très peu discutée. Pourtant, l'analyse chémoinformatique de la collection initiale de composés DEL pourrait fournir des informations utiles sur l'espace chimique couvert par une DEL particulière, sa diversité et son taux de réussite pour la découverte des molécules-candidates de futurs médicaments. Par conséquent, cette thèse de doctorat est axée sur la génération par ordinateur de milliers de DELs et sur l'analyse de leur espace chimique à l'aide de la Cartographie Topographique Générative[8] (*Generative Topographic Mapping*, GTM).

L'idée de la GTM consiste à insérer une hypersurface rectangulaire (appelé manifold) dans l'espace multidimensionnel défini par les descripteurs moléculaires, où le manifold passe aussi près que possible des zones les plus denses du nuage de données. Après avoir trouvé la forme optimale qui décrit les données, les molécules (représentées par des points dans cet espace) sont projetées sur le manifold. Puis ce dernier se déplie vers la forme plane ce qui résulte en une carte 2D visualisant l'espace chimique en question.

La carte GTM peut ensuite être colorée par différentes propriétés moléculaires, ce qui donne lieu à différents types de cartes. Par exemple, la carte peut être colorée par les valeurs de logP des molécules dans l'espace chimique (voir ***Figure 1***). Dans cette thèse la GTM est utilisée pour la visualisation de l'espace chimique des DELs en raison de sa haute performance pour l'analyse de grands volumes de données[9].

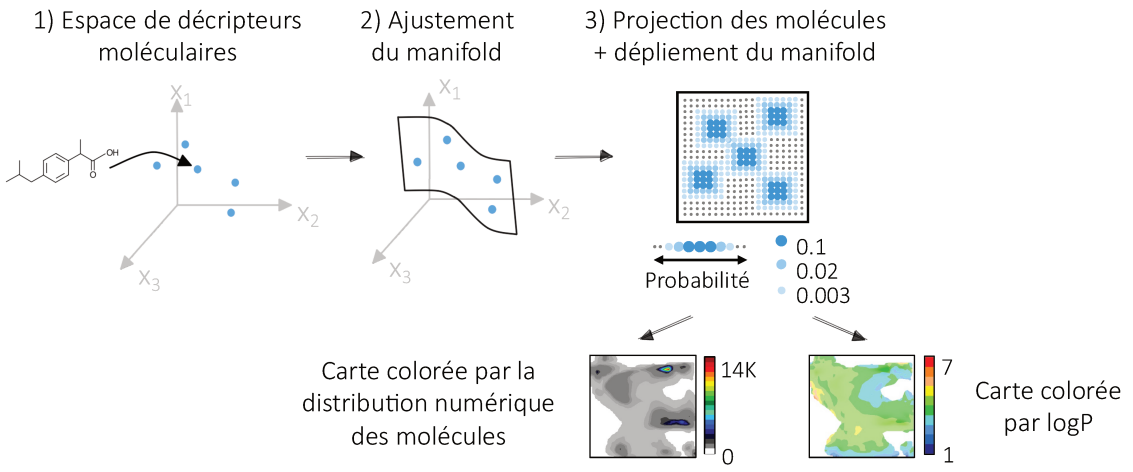


Figure 1. Méthodologie de la GTM : 1) les molécules sont représentées dans l'espace des descripteurs, 2) le manifold est inséré dans l'espace des descripteurs où il s'ajuste aussi près que possible aux données, 3) projection des molécules sur le manifold et son dépliement vers la forme 2D ce qui donne la carte de l'espace chimique, 4) chaque nœud de la carte est coloré soit par distribution numérique des molécules ou par une moyenne pondérée des propriétés des molécules se trouvant dans le nœud.

1.2 Résultats et discussions

1.2.1 Exploration de l'espace chimique de DELs et sélection d'une chimiothèque optimale pour le criblage primaire

Ce projet est consacré à la génération des DELs virtuelles et à l'estimation de leur pertinence pour le criblage biologique primaire lorsque peu ou pas d'informations sur une cible biologique et ses ligands sont disponibles[10]. 2.5K DELs contenant environ $2.5 \cdot 10^9$ composés ont été conçues en utilisant eDesigner[11], qui est un outil librement disponible pour la génération de DELs. Les DELs résultantes ont été analysées et comparées à la base de données ChEMBL[12] de molécules biologiquement testées (voir ***Figure 2*** pour le schéma détaillant le processus de génération et de comparaison). Cette dernière a été choisie comme chimiothèque de référence pour identifier la DEL optimale pour le criblage primaire en raison de sa diversité chimique et fonctionnelle – elle contient presque deux millions de composés testés contre plus de 15 000 cibles biologiques (version de ChEMBL28).

La comparaison des espaces chimiques des DELs à celui de ChEMBL a été effectuée à l'aide des cartes GTM visualisant l'espace chimique pour chaque chimiothèque. Par contre, étant donnée la subjectivité de la comparaison visuelle des cartes et le grand nombre de DELs, une métrique qui exprime l'intersection des espaces chimiques des deux chimiothèques sur les cartes GTM a été dérivée. De cette manière, l'ensemble des DELs a été classé en fonction de leur similarité structurale par rapport à la ChEMBL, en identifiant la DEL optimale contenant le pourcentage maximal possible de chémotypes biologiquement pertinents pour le criblage primaire. Des ensembles de trois et cinq DELs qui permettent d'atteindre une similarité encore plus haute par rapport à ChEMBL ont été identifiées et également proposées pour le criblage primaire.

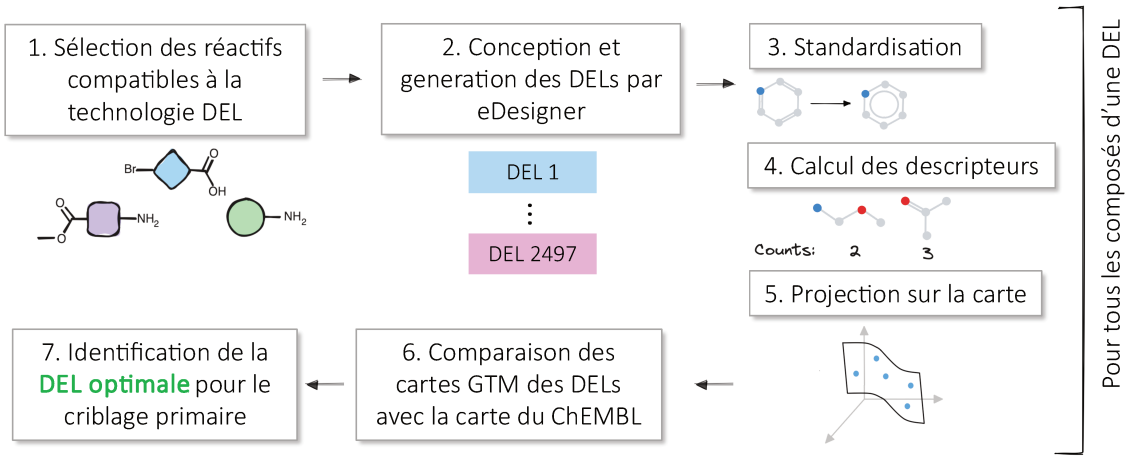


Figure 2. Processus de la génération de l'espace des DELs et comparaison avec ChEMBL.

1.2.2 Espace vectoriel des chimiothèques

Ce projet est consacré à l'analyse de l'espace de chimiothèques, donc l'espace où chaque DEL est considérée comme un objet individuel[13]. Cela est indispensable dans le cas d'analyse des DELs, car contrairement aux chimiothèques de composés classiques, la composition de la DEL ne peut pas être modifiée une fois synthétisée[4]. Cela signifie que la DEL doit être considérée comme un objet chémoinformatique autonome représenté à la fois comme une collection de molécules indépendantes et comme une entité individuelle[4]. De plus, étant donné qu'il est possible de concevoir plusieurs DELs différentes à l'aide de la chimie combinatoire, des milliers de DELs peuvent être générées et donc une méthode efficace permettant de comparer un grand nombre de chimiothèques est indispensable.

Pour cela, nous avons proposé quatre représentations alternatives de chimiothèques obtenues à l'aide de la GTM, définissant formellement l'espace de chimiothèques (voir ***Figure 3***). Ces représentations ne sont rien d'autre que les vecteurs codant les différents « paysages » obtenus en projetant une chimiothèque sur une carte : les « motifs cartographiques » de la chimiothèque (paysages de densité ou de distribution de propriétés). A l'aide de ces représentations vectorielles, on peut facilement et rapidement comparer et calculer la similarité structurelle et par propriété des milliers de chimiothèques ultra-larges.

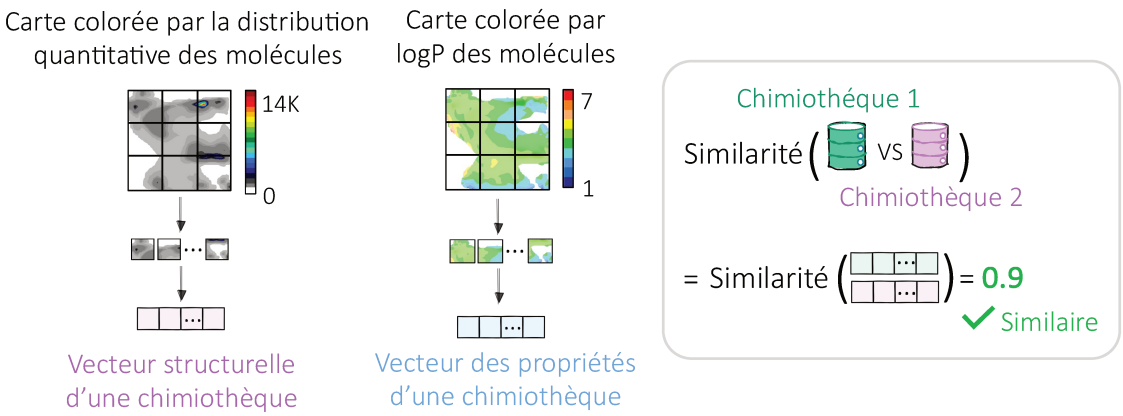


Figure 3. La création des représentations vectorielles pour les chimiothèques de composés à partir des cartes GTM.

A l’aide de la GTM, nous avons calculé la similarité des 2.5K DELs à la chimiothèque ChEMBL. Cela a permis de ranger de manière exhaustive les DEL en fonction de leur similarité structurelle et par propriétés (telles que logP, nombre de donneurs de liaisons hydrogène, etc.) avec ChEMBL et de prouver ainsi l'utilité de chacune des représentations vectorielles pour la comparaison des chimiothèques.

Du côté des applications, ces représentations peuvent être utilisées par les chimistes médicinaux pour choisir rapidement des chimiothèques pour la synthèse et les tests biologiques pour des projets spécifiques de découverte de médicaments, pour la diversification ou encore pour trouver des chimiothèques analogues, étant donnée la base de données de référence pertinente.

1.2.3 Meta-GTM : visualisation de l'ensemble des chimiothèques

Dans ce projet, nous proposons une visualisation de l'espace des chimiothèques en utilisant la meta-GTM ou μGTM[14,15]. Dans le contexte d'un large ensemble de 2.5K DELs virtuelles et de la base de données ChEMBL (utilisée comme référence), la visualisation de l'espace des chimiothèques (défini par les vecteurs des motifs cartographiques) est une manière intuitive d'obtenir une vue globale d'ensemble des chimiothèques diverses et de leur similarité. Cette vue d'ensemble peut également être étendue pour inclure toute autre chimiothèque, combinatoire ou non, afin de les localiser sur les cartes existantes.

Plusieurs μGTMs ont été créées, utilisant une optimisation paramétrique évolutive de la carte visant à préserver les distances inter-chimiothèques provenant de l'espace de chimiothèques initial sur la carte μGTM (voir ***Figure 4***). Ces cartes ont fourni un positionnement judicieux des DELs par rapport à la ChEMBL et les unes par rapport aux autres sur la carte, correspondant à leur similarité observée dans l'espace des chimiothèques initial défini par les différentes représentations vectorielles introduites précédemment.

La μGTM représente donc un outil efficace et utile pour :

(1) Fournir une vue globale de l'espace des chimiothèques et simplifier l'analyse des relations entre elles ;

(2) L'analyse de cet espace sous différents angles, positionnant les chimiothèques soit par similarité d'espace chimique soit par similarité de distribution de propriétés ;

(3) La sélection d'une chimiothèque de composés couvrant l'espace chimique et l'espace des propriétés désirées parmi des milliers de possibilités, en utilisant la base de données de référence appropriée.

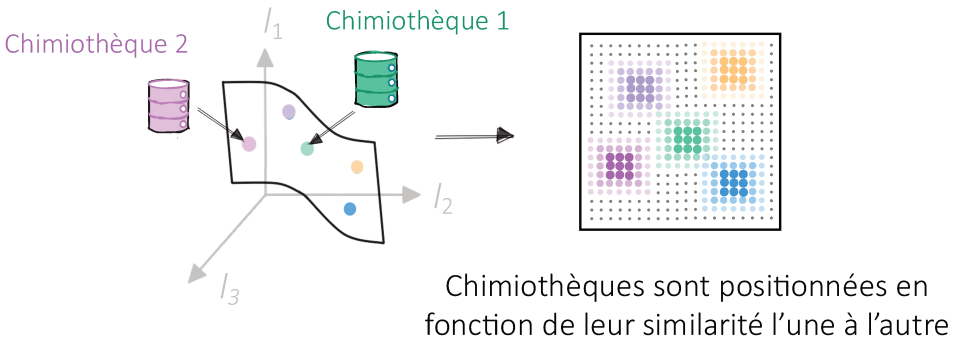


Figure 4. Visualisation de l'espace des chimiothèques sur une carte meta-GTM.

1.2.4 Analyse de la DEL ciblée contre la protéine BRD4

Dans ce projet, l'analyse de la DEL ciblée contre la protéine BRD4 a été réalisé en collaboration avec la société Novalix. Etant données les structures des réactifs à partir desquels la DEL a été synthétisée et les labels exprimant le rendement de la réaction (valide/invalide), des modèles de classification prédisant ce label ont été entraînés. Plus précisément, cela a été fait avec 153 différents ensembles de descripteurs structuraux ISIDA[16] en utilisant la méthode SVM (*Support Vector Machine*), soit linéaire, soit avec le noyau radial. Cela a donné lieu à 306 combinaisons descripteurs/noyaux. Ensuite, pour chaque combinaison, une validation croisée *5-fold* a été effectuée, résultant ainsi en un total de 1530 modèles individuels. Parmi eux, les modèles ayant atteint une précision équilibrée (*Balanced Accuracy, BA*) supérieure ou égale à 0,9 lors de la validation croisée sur l'ensemble de test ont été sélectionnés pour constituer le modèle de consensus.

Par la suite, l'analyse de l'espace de cette DEL à l'aide de la GTM a été faite et la comparaison de son espace chimique à l'espace des inhibiteurs de BRD4 provenant de la base de données publique ChEMBL28 a été réalisée. Cela a permis de voir la superposition entre plusieurs régions de l'espace des inhibiteurs déjà existants provenant de ChEMBL.

Finalement, les modèles de classification et de régression ont été entraînés en utilisant les données publiques sur l'activité biologique contre la protéine BRD4, disponibles dans la base ChEMBL, afin de prédire l'affinité des molécules présentes dans cette DEL. Ces modèles en combinaison avec les prédictions de l'activité faites par GTM ont permis de prioriser les molécules avec l'activité biologique la plus optimale pour la resynthèse hors ADN.

1.2.5 Visualisation de l'espace d'une chimiothèque combinatoire sur GTM sans énumération explicite des structures à l'aide de l'apprentissage profond

En règle générale, pour la visualisation de l'espace chimique d'une chimiothèque combinatoire (la génération de son « motif cartographique ») il faut :

(1) Enumérer les composés ;

(2) Effectuer leur standardisation ;

(3) Calculer les descripteurs pertinents pour la tâche en question ;

(4) Utiliser une méthode de réduction de dimensionalité (par exemple, la GTM).

Cependant, pour les chimiothèques combinatoires comme les DELs dont la taille peut aller jusqu'à 10^{12} molécules, ces calculs peuvent durer plusieurs jours. À titre d'exemple, le processus de génération de 10 cartes GTM pour 10 DELs, chacune contenant 10^6 molécules, dure 1 jour 10 minutes en utilisant la machine à 48 cœurs CPU (en cas de descripteurs structuraux ISIDA).

Afin de dépasser ces étapes longues, nous avons développé un réseau de neurones à convolution de graphe CoLiNN (*Combinatorial Library Neural Network*). Il permet de prédire la projection des composés combinatoires sur une carte GTM à partir des réactifs des composés en question ainsi que les indices des réactions nécessaires pour les énumérer (voir schéma sur ***Figure 5***). Cela permet de considérablement réduire le temps des calculs ainsi qu'éviter le stockage des représentations de structures énumérées, descripteurs, et les projections. Le modèle CoLiNN a été entraîné sur 388 DELs basées sur des schémas de réactions différents. Ce modèle a pu prédire avec une haute précision les cartes GTM des DELs non présentes dans l'ensemble de données d'entraînement.

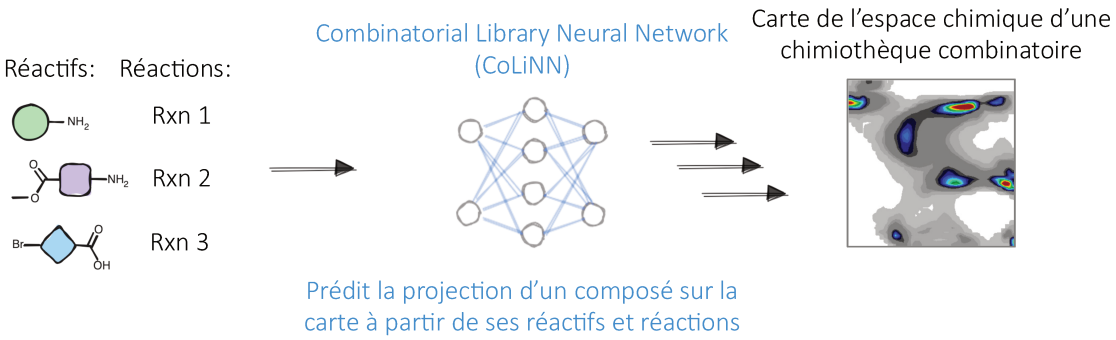


Figure 5. Visualisations de l'espace d'une chimiothèque combinatoire sans énumération des structures à l'aide du réseau de neurones à convolution de graphes CoLiNN.

1.3 Conclusion générale

Dans cette thèse, 2.5K de chimiothèques à codage ADN (DELs) ont été énumérées et analysées de manière exhaustive par leur composition structurelle ainsi que par distribution des propriétés moléculaires et physico-chimiques. Des nouveaux concepts et méthodes d'analyse et de comparaison par similarité des espaces chimiques combinatoires ultra-larges basés sur la GTM ont été développées en tenant compte des exigences particulières de ce type de chimiothèques (taille ultra-large, impossibilité de séparer les composés, grand nombre de DELs pouvant être conçues).

Parmi elles :

(1) Des métriques permettant la comparaison rapide et facile par structure des espaces chimiques ont été développées (en utilisant les informations provenant des cartes GTM) ;

(2) Le concept de l'espace de chimiothèques a été introduit et plusieurs représentations vectorielles (« motifs cartographiques ») des chimiothèques ont été proposées. Les métriques de similarité basées sur ces représentations ont été proposées pour comparer rapidement les milliers de chimiothèques par structure et propriétés;

(3) La méthode de visualisation de l'espace de chimiothèques (μGTM) où chacune d'elle est représentée comme un objet individuel sur la carte a été développée, ce qui facilite l'analyse de milliers de chimiothèques en une seule fois ;

(4) Des modèles d'apprentissage automatique ont été développés qui permettent de prédire :

- L'optimalité des réactifs pour la synthèse de DEL en terme de rendement chimique ;
- L'activité biologique contre une protéine spécifique.

(5) Un réseau de neurones à convolution de graphes (CoLiNN) a été créé pour la visualisation plus rapide et efficace de l'espace des chimiothèques combinatoires ultra-larges. CoLiNN évite l'énumération explicite des composés, standardisation, et calcul des déscripteurs moléculaires pour les composés d'une chimiothèque combinatoire en accélérant le processus d'analyse de son espace chimique.

1.4 Liste des présentations

Conférences nationales et internationales

1) **R. Pikalyova**, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, Application of molecular cartography to DNA-Encoded Library optimization, 11èmes Journées de la Société Française de Chémoinformatique, 5-6 Octobre 2023, Caen, France, communication orale et affiche

2) **R. Pikalyova**, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, Application of molecular cartography to DNA-Encoded Library optimization, Ecole d'été sur la conception de médicaments, 10-15 Septembre 2023, Vienne, Autiche, affiche

3) **R. Pikalyova**, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, Exploration of the Chemical Space of DNA-Encoded Libraries, 9ème atelier franco-japonais sur les méthodes computationnelles en chimie, 24-25 Avril 2023, Strasbourg, France, affiche

4) **R. Pikalyova**, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, Exploration of the Chemical Space of DNA-Encoded Libraries, La 8ème Ecole internationale d'été de Strasbourg en chémoinformatique, 27 Juin – 1 Juillet 2022, Strasbourg, France, affiche

Autres présentations

1) **R. Pikalyova**, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, Chemical Library Space: definition and DNA-Encoded Library comparison study case, Journée d'UMR 7140, 28 Avril 2023, Strasbourg, France, communication orale

1.5 Liste des publications

1) **Pikalyova R.**; Zabolotna Y.; Volochnyuk D. M.; Horvath D.; Marcou G.; Varnek, A. Exploration of the Chemical Space of DNA-encoded Libraries. *Mol Inform* **2022**, *41* (6), 2100289.

2) **Pikalyova R.**; Zabolotna Y.; Horvath D.; Marcou G.; Varnek A. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *J Chem Inf Model* **2023**, *63* (13), 4042–4055.

3) **Pikalyova R.**; Zabolotna Y.; Horvath, D.; Marcou G.; Varnek A. Meta-GTM: Visualization and Analysis of the Chemical Library Space. *J Chem Inf Model* **2023**, *63* (17), 5571–5582.

4) **Pikalyova R.**; Akhmetshin T.; Horvath D.; Varnek A. CoLiNN: A Tool for Fast Chemical Space Visualization of DNA-Encoded Libraries Without Enumeration. *ChemRxiv*. **2024**, doi:10.26434/chemrxiv-2024-qh3bn

2. Introduction

2.1 Need for Alternative Compound Screening Technologies

Despite the continuous progress in disease characterization, target validation, and medicinal chemistry, drug discovery remains a challenging endeavor. The scope of biological targets investigated in pharmaceutical research is expanding and many of them are classified as "undruggable" due to the disadvantageous screening metrics[17]. Drug resistance continues to be the obstacle for curing many patients from infectious diseases[18], cancer[19], chronic diseases such as epilepsy[20], inflammatory bowel diseases[21], etc. Optimization of drug properties such as safety, ADME-Tox profile, and selectivity is necessary since they drive the final clinical success of a drug-candidate[22]. This demands to search for more, better, and safer drugs[4,5].

High-Throughput Screening[3] (HTS) technology quickly became a primary method of identification of new chemical matter through screening of large compound libraries against targets of interest. The screening of HTS collections is based on a one compound-one well approach necessitating considerable investments in appropriate robotic equipment, material, and human resources. HTS compound libraries do not usually exceed the size of 10^6 compounds, limiting it to sampling only a small fraction of the theoretically available chemical space that is estimated to contain 10^{33} compounds[23]. HTS campaigns are based on biochemical/biophysical screening where a compound is screened in excess relative to the protein leading to challenges such as interferences due to compound aggregation and problems with analytical readout[4]. While in many cases HTS is fruitful, the associated challenges can sometimes lead to unproductive screening campaigns. Hence, there is a constant need for alternative technologies to address them.

DNA-Encoded Library (DEL) Technology[1,2] represents a complementary approach for hit identification offering many advantages compared to conventional screening methods. DEL technology consists in the creation of a usually ultra-large library of DNA-encoded compounds using water-based combinatorial chemistry and their simultaneous screening against a soluble biological target using binding affinity selection. DNA-encoded compounds are molecules labeled with single or double-stranded DNA. The latter plays the role of a "barcode" that encodes information about the building blocks (BBs) from which the compounds were synthesized[4]. This DNA barcode allows to easily

identify successful ligands bound to the protein after affinity selection via sequencing. DELs are usually synthesized using a combinatorial split-and-pool approach allowing the production of chemically diverse libraries of tens of millions to trillions of molecules[24]. DEL compounds are screened all at once in a single Eppendorf tube in contrast to individual compound screening in HTS. A simple experimental setup of affinity selection accessible both in industry and university laboratories allows cheap and fast hit identification.[4]

Despite DEL technology being here around for over 30 years now[1,4], the field of drug development has only just begun to give it the recognition it deserves. Pharmaceutical and biotech companies including GSK[4,25], X-Chem[4,26,27], WuXi AppTec[28], Amgen[4,29,30] (acquired Nuevolution), Eli Lilly[11], DyNabind[4,31–34], Vipergen[4,32–34], Google Research Applied Science[27], etc., are making noteworthy advancements in DNA-Encoded Chemistry for drug discovery as well as developing chemoinformatics platforms for DEL analysis[4,26,27]. Many success stories of employing this technology have been published, including DEL-derived hits that progressed to the clinic. According to Gironda-Martínez et al.[35], DEL-derived inhibitors of autotaxin (ENPP2) from X-Chem and of receptor-interacting protein 1 (RIP1) kinase and soluble epoxide hydrolase (sEH), both found by GSK, were in phase 1 and phase 2A of clinical trials, respectively, in 2021.

2.2 DNA-Encoded Library (DEL) Technology

DELs are usually prepared using split-and-pool combinatorial synthesis[36]. It consists of several steps, each followed by DNA enzymatic ligation reaction[4] (see **Figure 6**). First, a short piece of DNA (oligonucleotide of 7-15 base pairs long[37]) is covalently attached to a small molecule with an open functional group such as amino group[38]. This produces a so-called headpiece (see **Figure 6**), which can be chemically modified at the amine end and extended with the DNA tags via ligation at the oligonucleotide end[4]. The first oligonucleotide to ligate to the DNA of the headpiece is usually a primer which is a piece of DNA essential to initiate polymerase chain reaction (PCR). PCR is used in the compound identification step which is discussed at the end of this section. Once the primer was ligated, the solution containing the headpiece is split between different wells (**Figure 6**, a). Then, the first chemical building block (BB1) is added to each of the wells, where it undergoes a chemical reaction with the organic functional group of the headpiece (**Figure 6**, b, top). This is followed by the addition of the DNA tag encoding the first BB to the primer (**Figure 6**, b, bottom). The contents of all the wells are then pooled together (**Figure 6**, c) and the same procedure is repeated with the second BB (**Figure 6**, d, e, f). When the final synthesis cycle is finished, a closing primer is attached to the DNA tag of the last BB (not shown here for clarity). In the same way as the first primer, the closing primer is used in PCR initiation. The whole DEL is then stored in the Eppendorf tube.

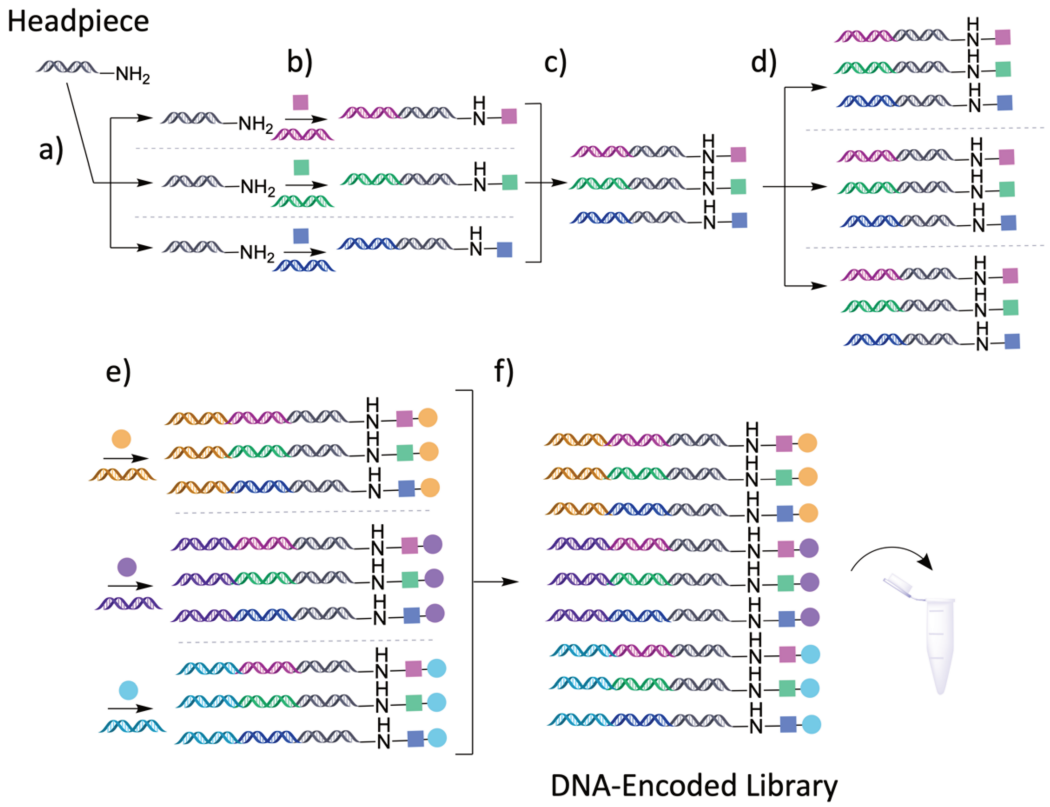


Figure 6. Split and pool synthesis of a two-building block DEL.

Screening of the DEL using affinity selection significantly differs from conventional chemical library screening in HTS. DEL compounds are mixed with the target protein immobilized on a solid support in a single vessel containing buffer solution as shown in **Figure 7**. All compounds compete for biological target binding. Molecules that failed to bind to the target protein are washed away with the buffered solution[4]. High-affinity binders, in their turn, are eluted from the protein typically by heating the solution which breaks up the interactions between them and partly denatures and detaches the immobilized proteins[4,5]. The DNA tags that encode high-affinity binders are further amplified using PCR technology and decoded by next-generation sequencing. Next, the processing of sequencing reads to counts of each barcode is done[39]. Counts are used to calculate the enrichment for each DEL member – ratio of counts in the presence of the protein / counts in the absence of the protein. Thus, it is considered that a higher enrichment value correlates positively with the affinity to the protein. In the next step, a list of compounds is selected to proceed in on- or off-DNA binding assay. The selection of this list can be based on multiple criteria including high enrichment values, counts, structural similarity to existent hits, and physicochemical profile[39].

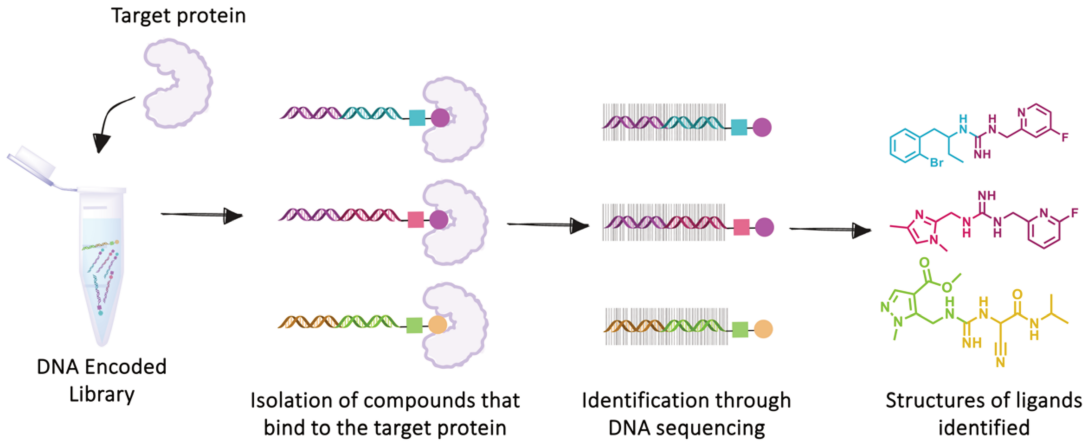


Figure 7. Scheme depicting the process of DEL affinity selection against a target protein and further identification of hits through DNA sequencing.

2.3 DEL chemoinformatics and its gaps

Over the past decade, the DNA-encoded library technology providing access to enormous chemical space became very popular both in academia and in the pharmaceutical industry[4]. However, early days of combinatorial and DEL chemistry proved that screening for affinity of large-sized DELs does not necessarily ensure success[40,41]. Without a proper design and analysis compound library's fate is predefined to failure – many such collections either did not improve the hit rate or displayed poor properties not sufficient for further medicinal chemistry follow-up[42]. Methods and tools for the design, enumeration, and analysis of ultra-large combinatorial compound libraries proved essential. This led to DEL chemoinformatics "gaining momentum" in the last decade with the development of DEL-adapted computational approaches of pre- and post-processing of DEL data[4,27,39].

2.3.1 Building Block selection

The first step in DEL design is the selection of suitable reagents for the library synthesis. Usually, either in-house BBs available in the pharmaceutical company are used or ordered from commercial suppliers. The reagents can be filtered by the BB class, number and nature of reactive sites, occurrence of structural alerts, DEL-compatibility (no DNA-intercalators, stable in water), presence of features important for binding (e.g. pharmacophoric), reagent price, etc. In addition, specific design guidelines to select novel reagents for drug-discovery purposes that were empirically verified can be followed. For example, the Rule of two (Ro2) filtering developed by Goldberg et al.[43] in AstraZeneca proved to be a success-driving way of BB prioritization for different drug discovery projects. However, it is useful not to include all available BBs in the production of DEL – clustering is commonly employed to remove BBs that are already abundantly represented[4]. Then, the actual acquisition of BBs for synthesis is carried out through careful selection of those with appropriate reactivity. This is usually performed experimentally by validating each reagent in a test reaction with the functionalized DNA conjugate [4]. Leveraging these guidelines can already provide a refined and high-quality list of BBs for DEL synthesis.

However, a closer look at the chemical space provided by a particular BB set would be particularly useful. It would allow to use only those BB combinations that help to advance to specific and/or novel chemical space regions. In a recent study by Paegel et

al.[44], the impact of different BB sets on the property distribution was shown by chemical space visualization using the UMAP dimensionality reduction method.

Therefore, in this thesis, the DEL space provided by commercially available BBs was investigated. A deep learning-based method allowing to take a look at the chemical space as a function of the BB set without explicit compound enumeration was developed and will be discussed in the following sections. Practical BB questions were addressed as well – reactivity label prediction QSPR models were developed, providing yet another way of BB filtering.

2.3.2 Enumeration

The combinatorial enumeration of compounds, using predefined reactions and available BBs, is a brute force approach allowing to fully explore the space of a combinatorial library. However, practically, compound enumeration is limited to the size of around 10^{12} compounds[4]. This limit is determined both by the efficiency of current enumeration algorithms and storage requirements. According to Goodnow and So's[4] estimation, for a library of 10^{12} compounds the storage space elevates already to 20 terabytes (if only 25 bytes per structure are taken). If enumeration is followed by a detailed chemical space or property distribution analysis, then, calculated descriptors and properties need to be stored additionally, increasing storage space requirements. The issue known as "combinatorial explosion" in combinatorial chemistry arises when the library size becomes so vast that it is no longer enumerable.

To address the combinatorial explosion problem, an eDesigner tool for efficient DEL design and enumeration was developed by Martin et al.[11] It does so by employing a structured and iterative approach to library generation, incorporating BB compatibility checks, representative sampling of the library, and efficient data handling. This ensures that only feasible and chemically relevant combinations are considered, significantly reducing the computational resources needed and allowing for the practical evaluation of vast DEL chemical spaces. Reactions encoded in eDesigner were experimentally validated by medicinal chemists from Eli Lilly making eDesigner a trustable tool that can be used in real DEL design tasks. In more detail, eDesigner relies on the following stages:

1. *Building Block Categorization:*
    - Identify and annotate functional groups in input building blocks.
    - Group building blocks into types based on functional group combinations.

- Eliminate incompatible combinations, which might lead to unwanted reactions or instability (e.g. strong electrophiles and nucleophiles are considered incompatible).

2. *Generation of Preliminary Compound Plans:*
   - Start with the functionalized headpiece and grow structures iteratively by adding compatible building blocks and reactions.
3. *Formation of Potential Libraries:*
   - Group compatible structures (which use the same reactions) into larger library designs.
   - Filter libraries based on the heavy atom count criteria in potential molecules selected by the user. In this step, molecules are not enumerated; the heavy atom count is estimated from its distribution in BBs and the average count of atoms gained or lost from a particular reaction.
4. *Enumeration:*
   - Generate a set of instructions for library enumeration that will be passed to LillyMol enumeration software.
   - Randomly sample a representative subset of the full library. Martin et al. reported that randomly sampled 10 000 compounds proved to be an optimal size allowing to achieve virtually indistinguishable property distributions from larger sizes during profiling experiments.

In such a way, eDesigner partially avoids the "combinatorial explosion" problem by generating only feasible library designs and sampling representative subsets instead of the full collection. Practically, however, when enumerating beyond 10M compounds in this thesis, eDesigner exhibited a significant slowdown, particularly towards the end of the process when nearly all library members had been enumerated. This presents a notable drawback from the user perspective, as the enumeration of larger libraries is not optimized, performed on a single CPU without parallelization, and can take days for tens of millions of compounds. Also, sometimes, it is worth analyzing the full library, especially if it is highly structurally diverse. These issues can be addressed either by optimizing the eDesigner software or by bypassing compound enumeration altogether.

eDesigner is the first and only DEL design software that is publicly available. Hence, in this thesis, eDesigner was used to create a DEL space containing 2.5K libraries using

commercially available BBs and encoded reactions inside the tool. Per each library, a 1M representative subset was enumerated resulting in 2.5B compound space overall. The actual full library sizes spanned from 1M to 7B compounds.

2.3.3 DEL chemical space analysis and comparison

The concept of the "chemical space" can be seen as either the discrete collection of all conceivable molecules or the multi-dimensional descriptor space that contains all potential molecules[45]. Its "exploration" or "navigation" can lead to discoveries of novel compounds with properties interesting for drug design. Therefore, advanced techniques for its analysis and visualization are essential to comprehensively understand its scope. Map-based methods are among the most popular visualization approaches, where the multi-dimensional descriptor space is mapped into a 2D plot, essentially, performing dimensionality reduction. Examples of such methods include Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Self-Organizing Map (SOM), and Generative Topographic Mapping (GTM), among others.

Chemical space comparison has been an active area of research in chemoinformatics since one of the ways to assess and further improve the quality/diversity of the compound library is to compare it with a reference dataset that possesses desired properties or characteristics. Library comparison is important in the context of compound acquisition, e.g. for diversification of in-house collections[46] or in the case of focused screening. Comparison of virtual compound libraries can lead to the identification of the optimal collection based on a variety of criteria, e.g. high structural diversity, property profile, ease of synthesis of molecules, etc.

To analyze and compare ultra-large compound libraries like DELs, strategies for efficient analysis and comparison of chemical spaces are needed. Despite the relevance of DEL compound space exploration and comparison, there was only one published work that discussed its analysis. In this study, Kontijevskis et al.[30] suggested Reduced Complexity Molecular Frameworks (RCMF) as molecular descriptors. They performed a diversity analysis of the DEL space by analyzing different combinations of RCMFs of the BBs using a heatmap. Overall, in that study, the RCMF and Bemis-Murcko scaffold analysis was performed for four DELs of size 107-151M compounds. However, the analysis of only four DELs is not enough to make any conclusions about the relevance of the whole DEL technology for drug discovery. Moreover, RCMF encodes information

about the molecular framework, such as the types and sizes of rings, the lengths of linkers, and the angle information. This representation is therefore too general and might overlook critical details about the chemical structures, potentially missing important nuances in the chemical and biological properties of the compounds. This can influence the precision of DEL space diversity analysis and comparison, particularly in the context of drug discovery, where detailed structural information is imperative for understanding molecular behavior and interactions.

Consequently, this thesis is dedicated to the generation of the ultra-large DEL space, its detailed analysis using a robust and confirmed chemical space visualization method - GTM[9,47–52], along with the development of library comparison methods adapted to handle thousands of DELs.

2.3.4 Property analysis

Usually, property distribution analysis of a compound library is done by means of histograms or descriptive statistics. However, for ultra-large libraries, theoretical calculation of properties can be time-consuming. Hence, Goodnow and So[4], verified whether a random sampling of a small portion of the library will be representative enough of the whole collection and thus provide the same property distribution histogram. This study demonstrated that a randomly selected 0.001% sample of a 1.6 billion compound DEL yields a property histogram identical to that of the entire library. However, histograms plotted for many compound libraries are difficult to comprehensively compare and interpret. This is especially critical for DELs – comparison of the property span of thousands of virtual DELs is not trivial with common statistics-based methods. In addition, histograms do not provide any information about structural diversity inside the property bar, although molecules with similar properties can be structurally highly different. Hence, in this thesis, cartography-based property distribution comparison[53,54] will be used to be able to account both for structural and property space overlap between libraries.

2.3.5 DEL hit analysis

After DEL affinity selection, DNA-tag counts are used to calculate the enrichment and thus determine whether the molecule should be synthesized on- or off-DNA and tested in a separate binding assay. However, DEL counts are noisy, meaning that molecules with the highest enrichment values are not necessarily the best binders. Focus on only high-

enrichment molecules can thus lead to missing out on interesting compounds and helpful structure-activity relationship (SAR) information[39]. Moreover, it may happen that the selected hit list compounds may be difficult to synthesize and/or have poor solubility to carry out a binding assay[39]. To overcome this, Quantitative Structure Property/Activity Relationship (QSPR/QSAR) models can be trained on DEL selection data to predict the enrichment. McCloskey et al.[27] trained a classification model on DEL selection data and showed that the learned SAR from the DEL screen allows to select potentially active compounds from commercially available low-cost libraries. Lim et al.[39] extended this approach by training a regression model on count data to predict the enrichment.

Nevertheless, not only the enrichment should influence the prioritization of DEL hits. In fact, a low correlation between DNA sequence counts and dissociation constant K_d values of small molecules was found by Mannocci et al[4,55]. A half-maximal inhibitory concentration predicted by machine learning models can be a relevant filter to consider. For example, publicly available biologically tested molecules with IC50 values can be used for training. Learned from the available ligands' SAR, the model can be further used to prioritize hits after the DEL selection experiment or even before selection. Hence, in this thesis, a QSPR model for hit prioritization by pIC50 prediction was created and evaluated on experimental data from the focused DEL tested by Novalix.

2.3.6 Overview of the gaps in DEL chemoinformatics

The complex and massive data associated with DELs created a unique demand for DEL-compatible chemoinformatics methods. Even though DEL chemoinformatics only started to emerge in the last decade, it is steadily evolving to address the requirements of DEL data analysis. However, until now most efforts were focused on the analysis of the libraries of BBs or identified active compounds[38,40,56–58] . The chemical space covered by DELs remains underexplored due to its extreme vastness and the necessity to enumerate it for a full investigation. Only one paper reported the analysis of DEL space using Reduced Complexity Molecular Frameworks (RCMF) methodology[30]. However, this analysis was limited to only four DELs (>5 × 108 compounds).

Generating and analyzing a larger virtual chemical space of DELs is necessary to fully explore their novelty and drug discovery relevance. Hence, in this thesis, an ultra-large DEL space of 2.5K libraries containing in total 2.5B compounds was enumerated and analyzed. Multiple DEL-adapted chemoinformatic methods for the analysis and

comparison of such a high number of large-sized libraries were developed. For the implementation of chemoinformatic methods for DEL chemical space analysis in this thesis, several important factors differentiating DELs from other compound libraries were taken into account:

1) A DEL is synthesized and tested as a whole, meaning it cannot be cherry-picked and thus needs to be treated as a separate chemoinformatic object.
2) DELs can be extremely vast and different library designs can be created, requiring 'big data' compatible and robust computational methods allowing to process many libraries at a time.
3) The size of the DEL does not guarantee the drug discovery program's success, a directed library design should be adapted instead.
4) DEL analysis needs time- and resource-consuming compound enumeration and methods of avoiding it are of high interest.

2.4 Methods of comparison/analysis of chemical libraries

To analyze 2.5K DELs, overall containing 2.5B compounds, powerful and efficient 'big data' compatible methods of chemical library analysis and comparison are needed. There are many types of chemical library comparison methods: graph-based, map-based, vector-based, fingerprint-based, and fragment-based. A summary of such methods is given in **Table 1**, with one example per type. These methods were selected since they all provide a metric to quantify the similarity between a pair of libraries, which can significantly accelerate the analysis of thousands of compound collections.

In a graph-based method, proposed by Fourches et al.[59], an entire dataset of compounds is represented as a Dataset Graph (DG), also known as a Chemical Space Network (CSN). In this graph, each point represents a molecule, with its position defined in a molecular descriptor space. Points within a certain distance range are connected by edges, thus giving a graph structure. Such DGs can be compared by structural similarity to each other by calculating graph index, e.g. average vertex degree or Randic connectivity index[59]. However, the Randic index does not account for the specific structural properties of the molecules, only the degrees of the vertices. Two graphs with very different molecular structures could have similar Randic indices if their degree distributions are similar, making this method unreliable for accurate chemical library comparison by chemotypes.

Miranda-Quintana et al.[60] introduced an extended version of fingerprint-based similarity metrics for the comparison of multiple compounds simultaneously. Instead of pairwise comparisons, the proposed extended similarity indices allow to compare sets of compounds (fingerprints) simultaneously. The extended similarity metrics are designed to be computationally efficient, scaling linearly with the number of compounds O(N). However, this method is limited by the fingerprint size, the longer ones being logically more informative and less prone to bit collisions.

SpaceCompare introduced by Bellmann et al.[61] is a fingerprint-based method of chemical space comparison. However, it calculates fingerprints not for full molecules but for fragments, thus allowing to avoid compound enumeration. SpaceCompare calculates the overlap between a pair of combinatorial spaces by using Connected Subgraph Fingerprints[62] (fCSFP) to represent and compare chemical substructures. In SpaceCompare, the overlap calculation works by first eliminating fragments that cannot

contribute to the overlap (covering step), thereby reducing the number of candidate products. Then the overlap calculation is performed on this smaller set of products (combination step), with the success of this process depending on how well the fragments match between the two spaces. Once the fragments that are covered in both spaces are identified, SpaceCompare enumerates candidate products and determines the overlap in counts by comparing their unique SMILES representations. However, when the overlap between two spaces is too high so that it is nonenumerable with available resources, SpaceCompare will not be able to operate. As the authors note themselves, if, for example, Enamine REAL Space had 50% of common products with a hypothetical chemical space of comparable size, SpaceCompare would not be able to calculate the overlap. In their work, Enamine REAL[63], CHEMriya[64], and other combinatorial spaces were successfully compared using SpaceCompare because they do not have a lot of overlapping products. The lowest and the highest overlap counts were 2867 (between REAL, Knowledge[65], and GalaXi[66] spaces) and 38M (between REAL and GalaXi spaces)[61]. Such compound numbers are easily enumerable; therefore, their approach did not fail to calculate the overlap. In addition, SpaceCompare does not provide explainable visualizations or indicate which specific regions of the chemical space overlap. In their work, only one molecule was shown as an example compound from the overlap.

Unlike previously mentioned approaches, map-based methods of chemical space visualization not only provide intuitive chemical space maps but also allow for accurate comparison of libraries by their structural and property similarities. GTM[8] is a powerful and comprehensible dimensionality reduction method whose ability to accurately visualize various chemical spaces was extensively tested[9,46,47,49,50,67–69]. Its idea consists in inserting a 2D hyperplane (manifold) into the multidimensional descriptor space where it adapts to the data cloud formed by molecules of the dataset. When the optimal form is found, data points are projected to the manifold with node-specific probabilities, and then it is folded back to the 2D form. The latter represents a chemical space map that can be colored either by the quantitative distribution of compounds across the chemical space, by class of compounds, or by properties, giving rise to different GTM landscapes. Using GTM, a library can be described as either a map or a vector. The latter, proposed by Gaspar et al.[50], is a cumulative projection vector that indicates the likelihood of compounds from the library to be projected into specific nodes on the map. A pair of

libraries can be quickly and elegantly compared by calculating any similarity metric between their cumulative projection vectors, for example, widely used in chemoinformatics Tanimoto coefficient. GTM, thus, allows to quickly quantify similarity between entire libraries offering explainability at the same time. In fact, comparison by cumulative vectors can be imagined by simple map overlap of two libraries. Therefore, low or high similarity is directly explainable by investigating the maps of the corresponding libraries. If needed, compounds from map zones can be also extracted for even more detailed structural analysis.

Another map-based method for chemical space analysis and visualization that is particularly interesting to investigate here is Multi-dimensional Scaling (MDS). The interest lies in the successful attempt by Agrafiotis and Lobanov[70] to use the MDS for visualizing a chemical library without compound enumeration. In their work, they trained a fully connected Multi-Layer Perceptron (MLP) to predict the coordinates of combinatorial products on the MDS map using only the descriptors of the respective building blocks (BBs) they are composed of. This approach allowed them to accurately predict the map for a 2-BB combinatorial library of 90,000 compounds. However, the MLP developed in their study was not given any information on the reactions used to obtain the products. However, omitting reaction information can potentially distort the predicted position on the chemical space map of a product synthesized using various reactions.

There are many methods of chemical space analysis and comparison, and all of them are generally applicable but the choice can vary depending on the expected quality of library comparison and its interpretability. In this thesis, the goal was to analyze and compare 2.5K DELs demanding both ‘big data’ compatibility, comparison accuracy, and interpretability – to be able to explain the similarity between libraries. GTM is an intuitive method of chemical space visualization that results in easily interpretable 2D maps allowing to analyze different aspects of the complex chemical space. It has been widely used in chemoinformatics for chemical library comparison[46,67] and property prediction[48]. Hence, herein GTM was selected as one of the best methods to explore the DEL space from different perspectives, both from a structural and property point of view. This thesis also tackles the challenge of bypassing the compound enumeration step by developing an

enumeration-free approach to GTM visualization, extending its use for even nonenumerable chemical space analysis.

Table 1. Summary of methods used to analyze and compare compound libraries.

	Type	Method	Library representation	Metric for library comparison	Pros	Cons	Ref.
1	Graph-based	Chemical Space Networks (CSN)	A library is represented by a CSN where two nodes - individual compounds - are connected if the similarity between them is higher than a given threshold.	Connectivity index.	- Supports comparison and visualization	- Not "Big Data" compatible - Enumeration of compounds required	Fourches et al.[59]
2	Map-based, Vector-based	Generative Topographic Mapping (GTM)	A library is described by a cumulative vector of probabilities of compounds to fall into particular nodes of the map created by the GTM algorithm.	Any similarity metric.	- "Big data" compatible - Interpretable - Supports comparison and visualization	- Training is limited by a frame set size - Enumeration of compounds required	Gaspar et al.[15]
3	Fingerprint-based	Extended Similarity Indices	A compound library is represented as a collection of binary fingerprints.	Extended similarity index, such as an extended Jaccard-Tanimoto index. This index sums the bit coincidences and normalizes them to reflect the overall similarity	- "Big Data" compatible	- The level of detail in the fingerprint depends on its length - Enumeration of compounds required	Miranda-Quintana et al.[60]

	Type	Method	Library representation	Metric for library comparison	Pros	Cons	Ref.
				between the libraries.		- No visualization of the compound space	
4	Fragments' fingerprint-based	SpaceCompare	A topology graph represents the overall structure of the compound space. Nodes in this graph represent pools of reactants described by fragment-based Connected Subgraph Fingerprint (fCSFP), while edges represent the chemical bonds formed during reactions.	Overlap is calculated using fCSFPs of fragments and is expressed in product counts.	- Ultra "Big Data" compatible - Enumeration of the whole product space is not required	- If the actual overlap between the two spaces is too large (nonenumerable with available resources), SpaceCompare will not be able to operate - Overlap expressed in counts does not inform about intersected chemotypes - No visualization of the compound space	Bellmann et al.[61]

2.5 Generative Topographic Mapping (GTM) for chemical library analysis and comparison

2.5.1 Generative Topographic Mapping (GTM)

Generative Topographic Mapping (GTM) is an unsupervised non-linear dimensionality reduction method based on manifold learning. The main idea of the GTM is the projection of the high-dimensional descriptor space to a 2D latent space as shown in **Figure 8**. This is done by inserting a 2D hypersurface called manifold into the multidimensional descriptor space where its form is optimized to fit the shape of the data (ensemble of molecules). The manifold is a square grid of Radial Basis Functions (RBFs, represented by radially symmetric Gaussians) and associated with them nodes. The optimization of the manifold's form is achieved by adjusting its hyperparameters, including its flexibility and smoothness (determined by the number of RBFs, the RBF width factor σ, and the spacing of the RBFs), as well as the map size and the regularization coefficient. Once the optimal form of the manifold is found, data points are projected to it with node-specific probabilities called responsibilities (r_{ik} – the responsibility of the molecule i to be projected to the node k). The responsibilities of a compound are represented by real numbers, summing up to 1.0 over all nodes. In the final step, the manifold is flattened out to obtain a 2D map with the molecules projected on it.

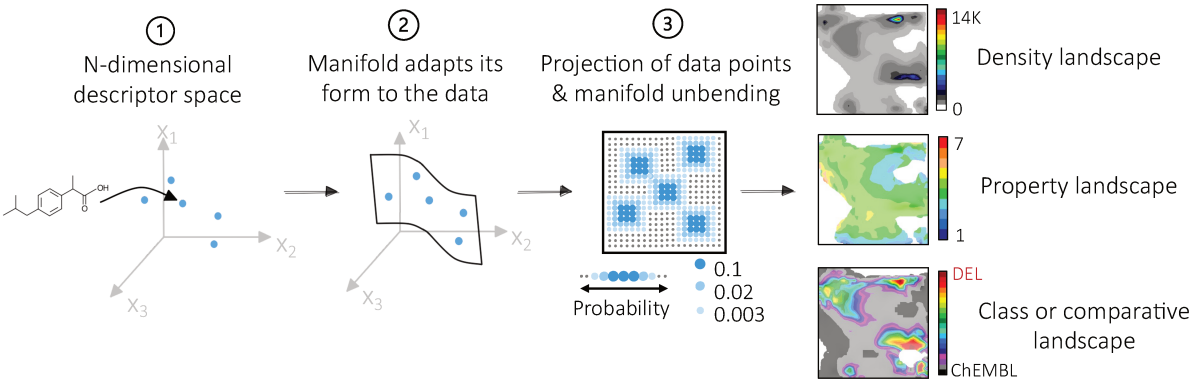


Figure 8. The concept of GTM. The data point (molecule) in the multi-dimensional descriptor space is projected on the manifold with node-specific probabilities called responsibilities. The 2D map is further colored either by the quantitative distribution of compounds giving a density landscape, by mean property value in each node giving rise to a property landscape, or by class. Here, the class is the belonging of the compound to a specific library - DEL or ChEMBL, resulting in a library comparative landscape.

2.5.2 Universal GTM

In this thesis, to visualize the chemical space of all compound libraries a 'universal' GTM, or UGTM[47], was used. UGTMs are maps trained to be poly-pharmacologically competent, namely, to be able to visualize a space of compounds that can be structurally and functionally diverse. In the study of Casciuc et al.[47], eight UGTMs were trained on ChEMBL23 data to be used as predictive models for the biological activity of molecules against 236 biological targets. In all research works presented in this thesis, apart from a focused DEL project, a UGTM1 built on ISIDA fragment descriptors (atom sequence counts with the length of 2−3 atoms labeled by CVFF force field types and formal charge status) was used.

2.5.3 Chemical Space (CS) and Chemical Library Space (CLS)

The conventional way of chemical library analysis consists in the investigation of its Compound Space (CS). The latter can be represented using molecular descriptors and visualized using GTM, resulting in a comprehensible 2D map. However, when combinatorial libraries like DELs are analyzed, an approach scalable to the analysis and comparison of thousands of libraries is needed. Moreover, a DEL is synthesized and tested as a whole, which makes cherry-picking impossible, meaning that it should be treated as a separate chemoinformatic object as well. Hence, in this thesis, a concept of Chemical Library Space (CLS) was introduced, where a compound library is represented as a numerical vector that encodes its structural information (see **Figure 9**).

Since GTM-produced maps of the chemical space are considered to preserve the topology of the initial multidimensional descriptor space, it is assumed that for a compound library:

1) Zones of the map are associated with its predominant chemotypes.
2) Cumulated responsibility (density) inside each node of the map reflects the chemotype distribution there.

Hence, in this thesis, several methods of vectorial compound library representation based on its GTM-produced map were proposed: different variations of Cumulated Responsibility Vectors or CRVs (Φ, Λ, and Ω) as well as Responsibility Pattern (RP) fingerprints (Γ) and RP count vectors (Γw). Regardless of the CRV type, it can generally be referred to as a CLS vector.

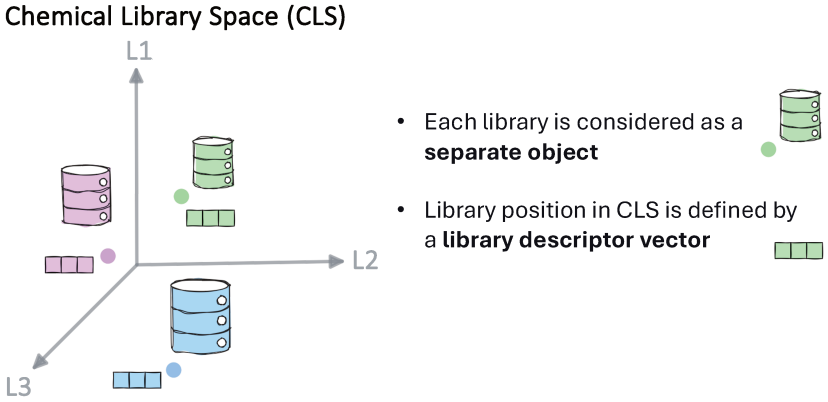


Figure 9. The concept of the Chemical Library Space (CLS) which is defined by library descriptors, here L1, L2, L3.

2.5.4 CLS vectors and GTM landscapes

Cumulated Responsibility Vector (CRV)

When multiple molecules of the chemical library are projected onto the GTM, a cumulative responsibility for each node *k*, namely, $\boldsymbol{c_k}$, can be calculated, as shown in **Equation (*1*)**. $\boldsymbol{c_k}$ is approximately equal to the number of compounds residing in node *k*. The $\boldsymbol{c_k}$ values can be rendered using a color code producing a map that visualizes the quantitative distribution of compounds across all nodes, which is called a density landscape (see **Figure 10**). It allows to spot the densest and scarcely populated areas of the chemical space of a single library. Alternatively, a *Cumulated Responsibility Vector*, ***CRV*** = $(c_1, c_2,\ldots,c_k)$ can be created and used to represent a whole compound library in the vectorial form. Its dimensionality is equal to K, the number of nodes on the map. This approach can be used to compare libraries to each other by calculating any similarity metric between the corresponding vectors.

$$\boldsymbol{c_k} = \sum_{i}^{N} \boldsymbol{r_{ik}} \qquad \textbf{(1)}$$

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

Normalized CRV (Φ)

However, ***CRV*** is library size dependent. Therefore, when libraries of different sizes are compared, $\boldsymbol{c_k}$ must be normalized by library size *N* according to **Equation (*2*)**. The

resulting normalized ***CRV*** or $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_k)$ encodes the compound distribution over the chemical space of the library irrespective of its size.

$$\boldsymbol{\phi_k} = \frac{\boldsymbol{c_k}}{\boldsymbol{N}} \quad \textbf{(2)}$$

Library-modulated CRV (Λ)

The ***CRV*** and $\boldsymbol{\Phi}$ treat all nodes of the chemical space map as equally important in describing the library. However, some nodes, particularly those highly populated by reference library compounds, may be more significant. To account for this, the ***CRV*** of the analyzed library (*a*) can be adjusted based on the compound distribution of a reference collection (*r*). This adjustment results in the library-modulated ***CRV*** or $\boldsymbol{\Lambda}=(\lambda_1, .., \lambda_k)$, which is calculated from the $\boldsymbol{\Phi}$ of both collections by determining the fraction of compounds from the analyzed library in the total population of each node, as shown in **Equation (*3*)**. In this $\boldsymbol{\Lambda}$ vector, $\boldsymbol{\Lambda_k}$ value of 0 is assigned to empty nodes that are not populated by any of the libraries. All others are assigned a value between $1 \leq \boldsymbol{\Lambda_k} \leq 2$ that varies as a function of the fraction of compounds of the library *a* in the given node. Nodes populated exclusively by compounds from either *a* or *r* have values of $\boldsymbol{\Lambda_k}$ of either 2 or 1, respectively. The fraction can be also rendered by a color code on the map showing zones populated by both libraries or predominantly by one of them. Such visualization is called a class or comparative landscape (see **Figure 10**). It allows one to spot unique and overlapping areas of the chemical space for a pair of libraries.

$$\boldsymbol{\Lambda_k} = \mathbf{1} + \frac{\boldsymbol{\Phi_k(a)}}{\boldsymbol{\Phi_k(a)} + \boldsymbol{\Phi_k(r)}} \quad \textbf{(3)}$$

$\boldsymbol{\Lambda_k}$ is the fraction of compounds of the analyzed library *a* in the total population of each node *k*

$\boldsymbol{\Phi_k(a)}$ and $\boldsymbol{\Phi_k(r)}$ are normalized cumulated responsibilities in the node k of the libraries *a* and *r*.

***Property-modulated CRV* (Ω)**

A library can be also analyzed in terms of its compounds' property distribution on a map. In this case, a mean property value per node ($\boldsymbol{\Omega_k}$) can be calculated according to **Equation (*4*)**. $\boldsymbol{\Omega_k}$ can be rendered by a color code on the map showing zones populated by compounds in a particular property range giving rise to a property landscape as shown in **Figure 10**. A property-modulated CRV or $\boldsymbol{\Omega} = (\sigma_1, \ldots, \sigma_k)$ can be calculated and used to compare compound libraries both by structural and property similarity simultaneously.

$$\boldsymbol{\Omega_k} = \frac{\sum_{i=1}^{N} \boldsymbol{P_i} \cdot \boldsymbol{r_{ik}}}{\boldsymbol{c_k}} \quad \textbf{(4)}$$

$\boldsymbol{\Omega_k}$ is the mean property value in the node *k* and $\boldsymbol{P_i}$ is the property value for the compound *i*

Library similarity score

To estimate the similarity between a pair of libraries, a pairwise Tanimoto coefficient can be calculated using their respective vectorial representations (Φ, Λ, or Ω):

$$\boldsymbol{Tc(a,r)} = \frac{\sum_k^K \boldsymbol{v_k(a) v_k(r)}}{\sum_k^K \boldsymbol{v_k}^2\boldsymbol{(a)} + \sum_k^K \boldsymbol{v_k}^2\boldsymbol{(r)} - \sum_k^K \boldsymbol{v_k(a) \cdot v_k(r)}} \quad \textbf{(5)}$$

Here, $\boldsymbol{v}$ is a chosen vectorial representation ($\boldsymbol{v}$ = Φ, Λ, Ω), and $\boldsymbol{K}$ is the total number of nodes.

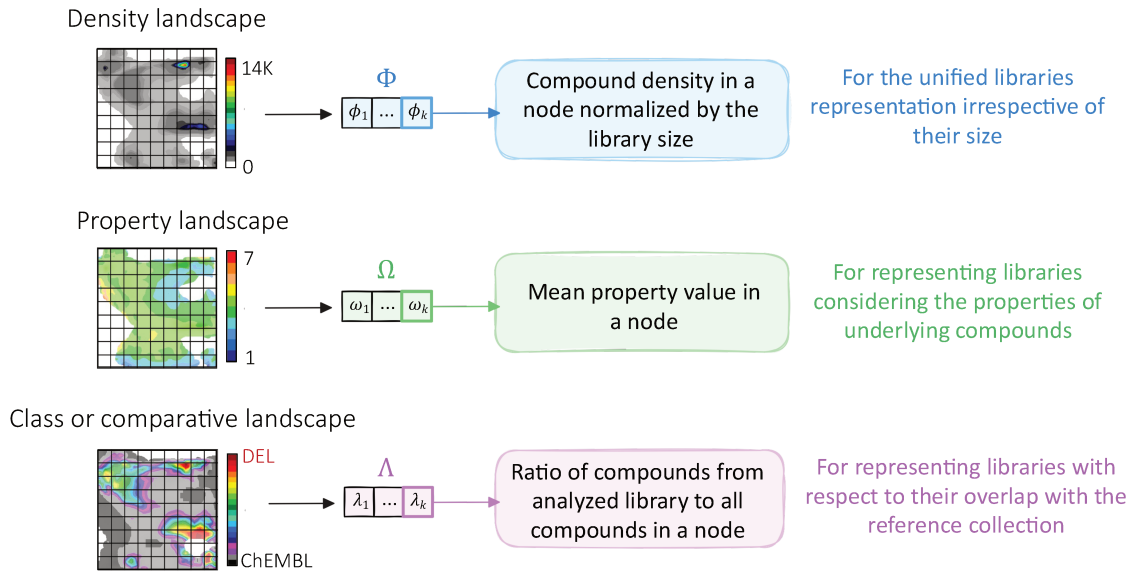


Figure 10. Scheme showing the creation of library descriptor vectors (Φ, Λ, and Ω) from different GTM landscapes of a compound library.

2.5.5 Responsibility Patterns

The position of a compound on a GTM is defined by a probability distribution over all nodes of the map, resulting in a responsibility vector R. For two structurally similar yet different compounds the R vectors will be very similar but not exactly the same (see **Figure 11**). Such compounds, that are projected onto similar zones of the map and thus having the same "chemotype" can be defined by the same Responsibility Pattern[71] (RP) vector that is more general than the R vector. The RP vector values for a molecule are calculated according to **Equation (*6*)**.

$$rp_{ik} = [10 \times r_{ik} + 0.9] \quad (6)$$

where [] is the truncation

rp_{ik} and r_{ik} are the RP and R values for a compound i in the node k

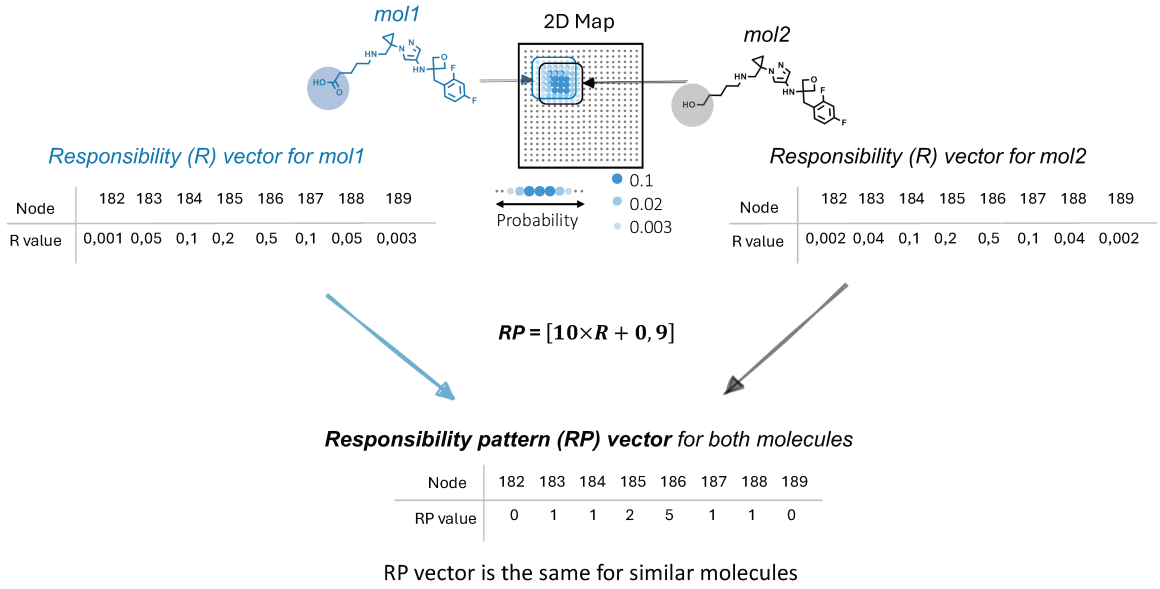


Figure 11. Example of the RP vector calculation from R vectors of two similar molecules (mol1 and mol2). The structurally different functional groups for the two compounds are underlined in circles. The R vectors of similar molecules give rise to the same RP vector.

Compounds that share an RP vector can exhibit varying degrees of structural similarity. They may share the same scaffold, Maximum Common Substructure (MCS), or pharmacophore, examples of molecules sharing the same RP vector are shown in **Figure 12**.

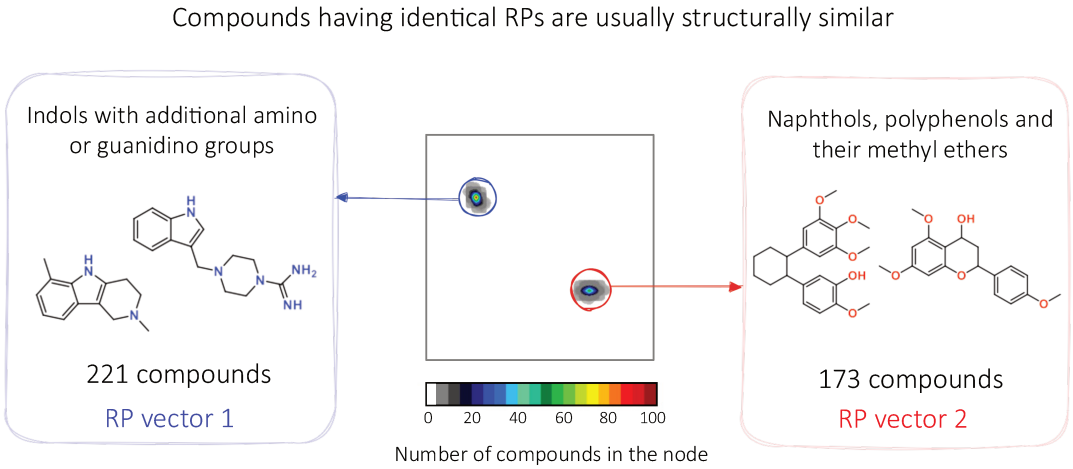


Figure 12. Structures of compounds coming from the same zones of the density landscape and thus sharing the RP vector.

To encode a compound library using RP, a library RP fingerprint (Γ) and RP count vector (Γ_w) were proposed in this thesis. Γ is a binary fingerprint encoding the presence or absence of a particular reference RP in the analyzed library, and Γ_w is a numerical vector with its values corresponding to the number of reference library compounds associated with a common RP present in both libraries. A scheme detailing the calculation of the Γ and Γ_w is given in **Figure 13**.

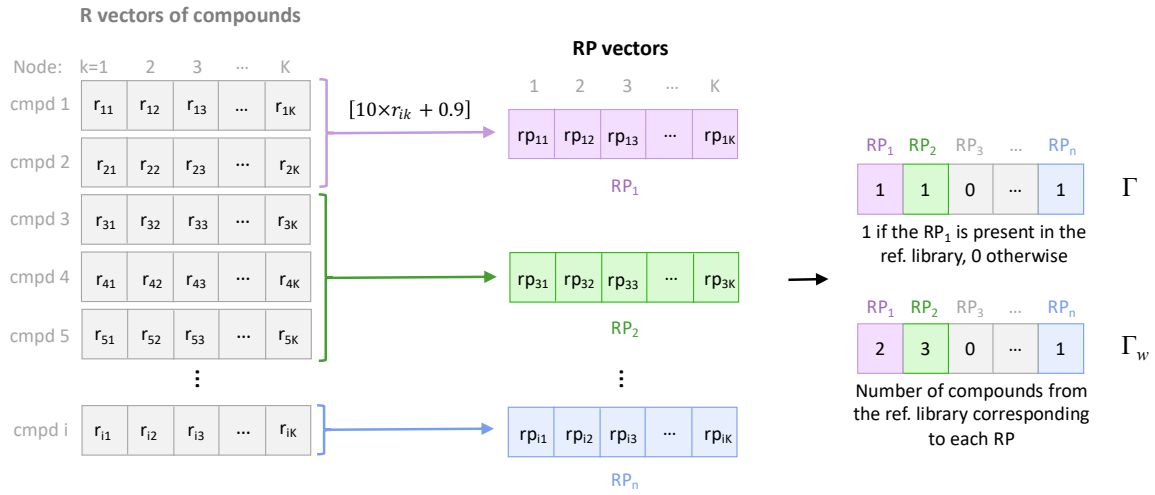


Figure 13. Scheme of the calculation of the Γ fingerprint and Γ_w vector for a compound library.

To compare two libraries by their RPs, coverage scores based on Γ fingerprint and Γ_w vector can be calculated. Coverage of a reference library (r) by a candidate library (a) can be defined in terms of Γ fingerprint overlap. For this, the fraction of RPs of a reference library also present in the analyzed library is calculated. In other words, in this case, the coverage score is the number of on-bits common for two libraries divided by the total number of on-bits in the reference library fingerprint, see **Equation (*7*)**.

$$\boldsymbol{Coverage(a,\ r)} = \frac{\sum_i \boldsymbol{\Gamma_i(a)\Gamma_i(r)}}{\sum_i \boldsymbol{\Gamma_i(r)}} \qquad \textbf{(7)}$$

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

Nevertheless, the Γ fingerprint accounts only for the presence or absence of a particular RP, not for the number of compounds associated with it. Since different RPs can be populated differently, high RP coverage does not imply high compound coverage.

To account for the number of compounds, a weighted RP coverage score can be calculated, see **Equation (*8*)**. It is defined as the fraction of compounds of a reference library that correspond to the common RPs (present in both analyzed and reference libraries).

$$\boldsymbol{wCoverage(a,\ r) = \frac{\sum_i \Gamma_{wi}(r)\Gamma_i(a)}{N_r}} \quad \textbf{(8)}$$

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.

2.5.6 Meta-GTM

CLS vectors are used to compare a pair of compound libraries. However, in some cases, it may be necessary to look at the relationship of three or more libraries at once to have a bird's eye view of the whole CLS. Therefore, in this thesis a meta-GTM approach[50], allowing to visualize the CLS composed of thousands of libraries from different perspectives was used to analyze the 2.5K DEL space. The main idea of the meta-GTM is to reduce the dimensionality of the high-dimensional CLS to give a 2D map where entire libraries are mapped objects. The meta prefix here is used to highlight the fact that the GTM is applied for the second time.

In this thesis, to create optimal meta-GTMs to visualize the DEL space relative to the ChEMBL reference database, map parameters were optimized using a Genetic Algorithm (GA). GAs is a class of optimization algorithms inspired by the principles of natural selection and genetics. They operate through the iterative process of selection, crossover, and mutation, which allow candidate solutions to evolve towards an optimal solution. In the case of map parameter optimization, the GA begins with the random generation of an initial population of map parameter combinations (number of map nodes, number of RBFs, RBF width, regularization coefficient value, preprocessing option of CLS). Each parameter combination is evaluated using a fitness function that quantifies the map visualization performance.

Meta-GTM fitness was defined by the quality of the preservation of interlibrary distances on it as calculated in the initial CLS descriptor space. The goal was to select a meta-map where the top 100 closest DELs to a reference dataset (ChEMBL) in the CLS

will stay in the top 100 closest neighbors of ChEMBL on the meta-map. For this, the list of the top 100 libraries was established by ranking 2.5K DELs by their Tc similarity to ChEMBL in the corresponding CLS. The Tc was calculated as shown in **Equation (*5*)**. Any DELs immediately beyond the 100th position but having the same Tc value as the 100th-ranked DEL were also included in the list, for consistency reasons.

After the manifold - defined using the current parameter set - is trained, ChEMBL and DELs are projected on it. The latent space coordinates (x, y) for each library are calculated as the geometric center of their responsibility clouds. Then, the Euclidean distance of each DEL relative to ChEMBL is calculated according to **Equation (*9*)**. Based on the ranking according to Euclidean distances and considering the 100 top DELs as "positives" and all other libraries as "negatives", a Receiver Operating Characteristic (ROC) can be calculated and plotted. The ROC Area Under the Curve (ROC AUC) is then calculated and used as a fitness function. If the top 100 DELs remain the closest 100 neighbors of ChEMBL on the meta-map, this will result in a high ROC AUC value and thus high performance of the map. The details of GA optimization duration and training are given in Chapter 6 (Meta-GTM: a tool for Chemical Library Space visualization).

$$\boldsymbol{D(DEL, ChEMBL)} = \sqrt{(\boldsymbol{x_{ChEMBL}} - \boldsymbol{x_{DEL}})^2 + (\boldsymbol{y_{ChEMBL}} - \boldsymbol{y_{DEL}})^2} \quad \textbf{(9)}$$

Different library descriptor vectors were used to represent DELs and ChEMBL resulting in Φ, Λ, or Ω-based CLS. Accordingly, for each CLS a meta-GTM was created. Meta-maps built on Φ and Λ library descriptors were used to analyze the CLS from a structural similarity point of view. Whereas Ω-based meta-maps allowed to analyze the CLS both from structural and property similarity perspectives. Five properties were selected for the CLS analysis: molecular weight (MW), logP, number of H-bond donors, number of H-bond acceptors, and quantitative estimate of drug-likeness (QED).

Meta-GTM landscapes allowing to color the CLS according to either the density of libraries (measure of how much the CLS zones are crowded) or any intrinsic library characteristic can be created. Libraries can be assigned a class, for example, as in the case of DELs, they can be either included or not in the top 100 neighbors of ChEMBL. The class of the library can be rendered using a color code on the meta-GTM landscape, facilitating the CLS navigation. Landscapes colored by library characteristics can be as

well useful. For example, they can be colored by the estimated price to synthesize the library, the reaction type, the number of BBs engaged in the library, library size, etc. This can significantly simplify the comparison of thousands of libraries based on custom criteria.

2.6 ML modeling for DEL hit prioritization and BB reactivity prediction

The two major steps involved in DEL technology – BB reactivity validation and hit prioritization are among the most challenging.

First, before purchasing BBs for DEL synthesis, their reactivity should be carefully validated experimentally. For this, each BB is combined with an appropriately functionalized BB partner attached to the DNA headpiece[4]. However, when multiple large-sized DELs need to be synthesized and if BBs are completely novel, such reactivity tests can require a lot of investment and time. Hence, in this thesis, to rationalize the validation process of reagents, the BB reactivity prediction problem using machine learning (ML) models was addressed.

Second, the identification of promising hit series can be difficult due to the noisy affinity selection results. Consequently, ML models that either denoise these data or guide the selection of the most promising compound series from the typically extensive hit list are highly preferred. This thesis investigates the latter - hit prioritization using ML model predictions of activity.

The Support Vector Machine (SVM) method was employed for this purpose, since itself and its extension – Support Vector Regression (SVR) are capable of resolving nonlinear Structure-Activity/Reactivity Relationships in the original descriptor space through the use of kernel functions, otherwise called "kernel trick"[72].

2.6.1 SVM and SVR

SVM is a robust supervised ML method used for compound classification tasks and its extension, SVR, is used for property prediction problems [72,73]. The idea of SVM is given the training data points, defined by a descriptor vector $x \in \mathbb{R}^D$ and their class labels $y \in \{-1, 1\}$, find a decision boundary the best separating two classes from each other (see the left side of **Figure 14**). This is done by projecting the data points into the descriptor space where SVM constructs a hyperplane that optimally separates the two classes. A hyperplane in an SVM is defined by **Equation (*10*)**.

$$H: \langle w, x \rangle + b = 0 \qquad \textbf{(10)}$$

x - represents a point in the feature space (a molecular descriptor vector).

w - is a weight vector perpendicular to the hyperplane.

$\langle w, x \rangle$ – denotes the dot product between w and x.

b - is the bias term that shifts the hyperplane.

Data points of one class that are the closest to the samples of another class are called support vectors (black overlined circles in **Figure 14**). They form positive and negative class hyperplanes H^+ and H^- that define the margin of the hyperplane [72,73]. Other SVM parameters include gamma γ and C parameters. The former controls the influence of individual data points: a low gamma value results in a smooth, generalized decision boundary by giving each point a large influence range, while a high gamma value creates a more complex, wavy boundary by giving each point a small influence range. The C parameter balances the trade-off between maximizing the margin and minimizing classification errors, with high C values focusing on correctly classifying all training data points (which can lead to overfitting) and low C values allowing some misclassifications to achieve a wider margin for better generalization. During the optimization process, the SVM algorithm adjusts the position and orientation of the hyperplane to maximize the margin between the classes while ensuring the best classification performance. This involves finding the optimal values for the parameters w and b that define the hyperplane H. In the same way, the parameters γ and C are selected so that there is a balance between margin maximization and classification accuracy. Once the optimal hyperplane is defined, test data are projected into the descriptor space and classified based on the side of the plane they fall on. In more detail, each data point x_i is classified by calculating the value of the decision function $f(x_i) = w \cdot x_i + b$; if $f(x_i) > 0$, the point is classified as positive (+1), otherwise as negative (-1).

In SVR, training data points are defined by a descriptor vector $x \in \mathbb{R}^D$ and their numerical target values $y \in \mathbb{R}$. SVR tries to find a function $f(x) = \langle w,x \rangle + b$ that best approximates the relationship between input features x and the target value y [72,73] (see right side of **Figure 14**). Small deviations between observed and predicted y values are allowed within the ϵ-insensitive tube (a margin), while errors greater than ϵ are penalized.

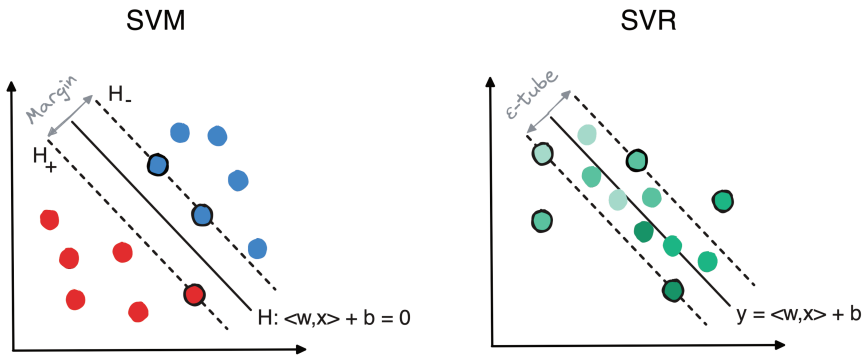


Figure 14. Schematic of SVM and SVR modeling.

In this work, both linear and RBF (Radial Basis Function) kernel SVM were used. The linear kernel is used for linearly separable data and is represented by the linear function (dot product) of the feature vector *x* and *w*. RBF SVM uses a non-linear function (RBF kernel) to map the input features into a higher-dimensional space where a linear separation might be possible (see **Figure 15**). RBF kernel, ($K_{RBF}(x_i, x_j)$), is a similarity function that is inversely proportional to the Euclidean distance between two points x_i, x_j. It replaces the dot product in the decision function, see **Equation (11)**.

$$K_{RBF}(x_i, x_j) = exp(-\Upsilon \parallel x_i - x_j \parallel^2) \quad \textbf{(11)}$$

$\Upsilon = \frac{1}{2\sigma^2}$

x_i, x_j - the feature vectors of two data points

$\parallel x_i - x_j \parallel$ - the Euclidean distance between x_i and x_j

σ – the parameter that controls the width of the Gaussian function used for mapping

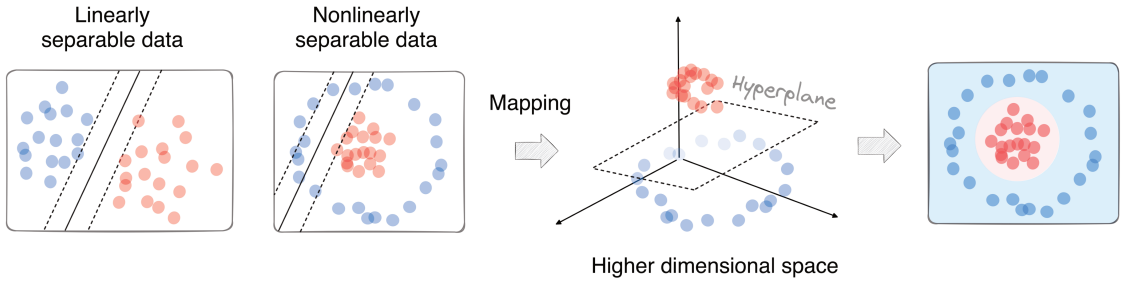


Figure 15. Scheme showing linearly and nonlinearly separable data. To find a decision boundary for the latter, the kernel trick using a non-linear mapping function is performed – this process is shown on the right of the figure.

3. Thesis outline

DEL technology has emerged as a powerful method of hit identification due to its ability to access the ultra-large chemical space while using a comparably simple experimental setup. Compared to conventional HTS screening, the hit identification process using DELs imposes unique demands on chemoinformatic methods. First, the screening for affinity of multi-billion-sized DELs did not prove to result in higher hit rates. It soon became clear that the design of the library, rather than its size, is the crucial factor. Consequently, there arose a demand for chemoinformatic approaches to facilitate intelligent library design through thorough building block analysis, chemical space exploration, and property distribution investigation. The practical challenges, such as the low interpretability of noisy affinity selection results hindering straightforward hit series identification, necessitate the use of machine learning models to denoise these data or guide hit prioritization. While DELs offer the advantage of exploring ultra-large chemical spaces, they also face the challenge of combinatorial explosion. The compound enumeration procedure limits the analysis to around 10^{12} compounds[4] and slows down the overall library analysis. Hence, approaches that skip compound enumeration are required.

Therefore, the main goal of this thesis was to develop efficient methodologies for the design and analysis of thousands of ultra-large DELs. The contributions and novelty of this thesis can be described as follows:

- *Generation of the ultra-large DEL space:* For the first time, a virtual space of 2.5K DELs was generated from commercially available building blocks using eDesigner, 1M molecules per DEL were enumerated resulting in 2.5B compounds.
- *Visualization of DEL space using GTM:* Each DEL's space was analyzed using GTM, providing a multi-perspective view of each library – its space was investigated both by structural diversity, overlap with publicly available databases, and property distribution.
- *Development of approaches for directed DEL selection:* A GTM-based approach for library selection based on the expected chemical space coverage

was proposed. This allowed to select an optimal structurally diverse DEL for primary screening.

- *Development of efficient chemical library comparison approaches:* Methods of efficient and quick pairwise comparison of DELs with a reference database were developed by introducing the Chemical Library Space (CLS) concept and library descriptors. Conventional similarity metrics calculated using such vectors showed to quickly and accurately rank DELs by their structural and property similarity to a reference database.
- *Application of the meta-GTM approach to Chemical Library Space visualization:* For the simultaneous analysis of 2.5K DELs by structure and property a meta-GTM approach was used[15]. It allowed to obtain a "bird's eye" view of the DEL space on one map.
- *Prediction of BB reactivity and hit prioritization for BRD4 focused DEL:* Affinity selection results of the focused DEL against BRD4 protein provided by Novalix were thoroughly analyzed. Both GTM- and QSAR-based approaches were applied for the analysis and prioritization of DEL hits. Furthermore, building block reactivity prediction models were developed offering a complementing rational way of their validation for synthesis.
- *Visualization of the DEL space without compound enumeration:* A highly performant deep learning model for the prediction of DEL compound position on the GTM without the need for compound enumeration was developed. This model will significantly reduce the time and resource consumption for the chemical space analysis of ultra-large combinatorial libraries.

4. Exploration of the chemical space of DNA-Encoded Libraries

Introduction

To systematically explore and understand the drug discovery potential of DELs, their chemical space should be comprehensively analyzed. However, most of the DEL-related information stays undisclosed inside pharmaceutical companies[4]. To our best knowledge, there was only one article analyzing DELs coming from Nuevolution company[30] (now Amgen). Nonetheless, in that study, the analysis was limited to only four libraries, and no structures were given. Hence, in this work, an ultra-large space of 2.5K DELs was designed using commercially available BBs, and 1M compounds per DEL were enumerated using the eDesigner tool resulting in 2.5B molecules in total (without DNA tags). Selecting a DEL suitable for a particular drug discovery project from such a large pool via exhaustive affinity screening is impractical. Therefore, here, we propose to use a GTM-based analysis of the chemical spaces of virtually generated DELs to select an optimal one for the drug discovery task in question.

Glossary

DEL – DNA-Encoded compound Library, which is usually an ultra-large combinatorial compound collection. The DEL compound is covalently attached to a DNA tag that encodes information about its building blocks. The whole DEL is screened simultaneously against a biological target in a test tube allowing to explore large chemical space regions at once. DNA tag is used for the identification of successful binders after DEL affinity screening.

eDesigner – A freely available tool for DEL design and enumeration. It is optimized to generate only feasible library designs and supports sampling of representative subsets instead of the full collection.

GTM – An efficient probabilistic dimensionality reduction method, compatible with "big data" analysis.

UGTM – A GTM that was trained to be "poly-pharmacologically competent" allowing it to accommodate ligands of different biological targets. Thus, it can be used to visualize structurally diverse chemical space containing biologically relevant chemotypes.

The GTM-based approach of library selection was applied to identify an optimal DEL for primary screening – when there is no or very little information about the biological target and its ligands. Such a library should be structurally and functionally diverse and contain biologically relevant chemotypes. ChEMBL is a publicly available

database of biologically tested compounds against >15 000 targets that displays such characteristics, thus being an optimal reference library. Therefore, each of the 2.5K DELs was compared to ChEMBL using UGTM. The comparison was accelerated by deriving a GTM-based chemotype coverage metric measuring how well a particular DEL covers the chemical space of ChEMBL.

Research Article www.molinf.com molecular informatics

doi.org/10.1002/minf.202100289

Exploration of the Chemical Space of DNA-encoded Libraries

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

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

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

1 Introduction

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

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

However, up to this point, most efforts were focused on the analysis of the libraries of BBs or identified active compounds.[3] Authors were less keen to explore the entire

[a] R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek
University of Strasbourg, Laboratory of chemoinformatics
4, rue B. Pascal, Strasbourg 67081, France
phone/fax: +33 368851560
E-mail: varnek@unistra.fr
[b] A. Varnek
Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University
Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan
[c] D. M. Volochnyuk
Institute of Organic Chemistry, National Academy of Sciences of Ukraine
Murmanska Street 5, Kyiv 02660, Ukraine
[d] D. M. Volochnyuk
Enamine Ltd.
78 Chervonotkatska str., 02660 Kyiv, Ukraine

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202100289

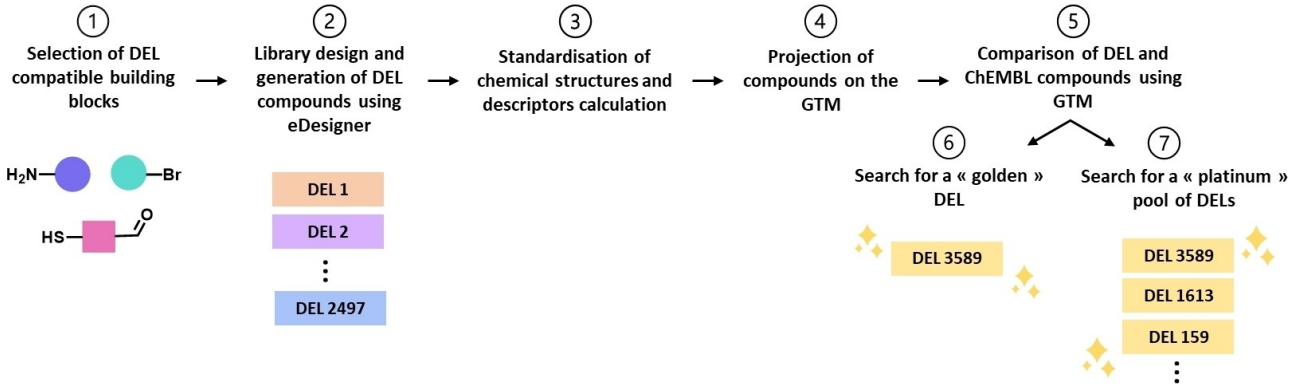


Figure 1. Workflow of the project. The rectangles represent separate DELs.

chemical space covered by DELs because it is extremely vast. To our best knowledge, only one paper reported the analysis of DEL space using Reduced Complexity Molecular Frameworks (RCMF) methodology.[11] However, in that work, the analysis was limited to only four DELs ($>5\times10^8$ compounds). Since DEL technology is actively being developed and new methodologies for DEL synthesis were being elaborated, the aforementioned pioneering work no longer reflects the status quo.

This work is focused on the generation of possible DELs from commercially available BBs using a tool for DELs generation called eDesigner.[12] Since screening thousands of DELs containing billions of compounds is unfeasible, we suggest choosing the so-called "golden" DEL(s) that covers the chemical space of biologically tested compounds to the highest extent. Such a library would have high structural diversity and contain the majority of biologically relevant chemotypes, which is critical for the success of the primary screening against novel biological targets. It was identified by comparing the generated DEL space to the chemical space of biologically relevant ChEMBL[13] compounds using Generative Topographic Mapping (GTM) – an efficient dimensionality reduction method.[14] GTM has proved to be a powerful tool for "Big Data" analysis and visualization (up to 1B compounds).[15] Notably, the prior development of quantitatively validated, polypharmacologically competent Universal Maps (uMaps) allowed us to propose a chemically meaningful representation of the to-date explored drug-like chemical space.[16] Only one of the several uMaps (uMap1, see the corresponding article) was used in this study for simplicity, but the study could be extended to consensus mapping on several uMaps.

2 Methods

2.1 General Workflow

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

2.2 Selection of Building Blocks

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

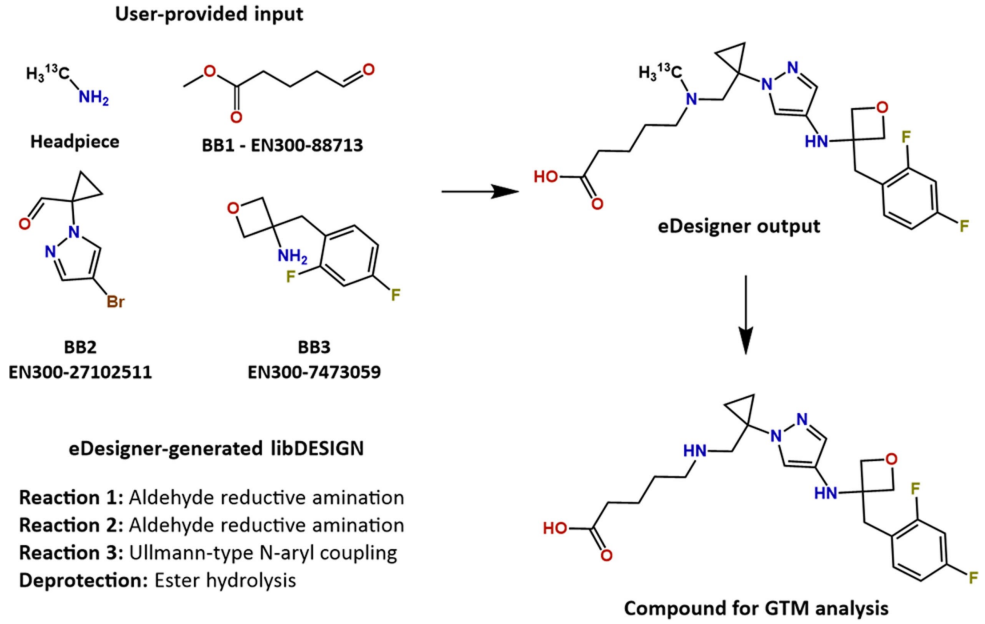


Figure 2. Example of DEL compound generation by eDesigner. The user should provide the headpiece and the list of BBs; an appropriate list of reactions will be selected automatically by eDesigner, and respective compounds will be generated. The isotopic mark is placed by eDesigner in order to know the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

2.3 DEL Generation with eDesigner

For the generation of chemical space of DELs, the eDesigner[12] tool was used (see Figure S5 in SI). At first, based on the list of the most efficient DNA-compatible reactions encoded in the tool (see Supporting Information of respective article[12]) and a user-provided list of BBs, it generates a special set of instructions for DEL compound enumeration called libDESIGNs. Each libDESIGN contains information about the starting headpiece (the whole DNA part for computational convenience is formally represented as a ^{13}C atom), the reaction types, and BBs which will be used in them, as well as deprotection reactions for the final stage of DEL generation. There are also several restrictions that can be applied to control some of the properties of the resulting DEL. They include, for example, the maximum and the median value of heavy atom count in the generated molecules, minimum library size, etc. Once the libDESIGNs are created, the representative DELs subsets of the selected size can be enumerated by the LillyMol tool.[21] An example of such enumeration is shown in Figure 2. The isotopic mark on the carbon atom specifies the place of attachment of the DNA tag. For clarity reasons, before physicochemical properties calculation and GTM analysis, the ^{13}C atom is removed, therewith obtaining the compound that would have been resynthesized off-DNA for validation in case of being selected during a real screening campaign.

2.4 Generative Topographic Mapping (GTM)

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

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

manifold is flattened out to obtain a 2D representation of the map with compounds projected onto it.

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

2.5 Universal GTM

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

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

2.6 Responsibility Patterns

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

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

2.7 ChEMBL Coverage Estimation

First, RPs for all compounds are calculated as described above. Then the pairwise overlap between each DEL and ChEMBL ($Ch\_RPs\ cov\ \%$) is determined by dividing the number of common RPs for both libraries ($N_{common}(Ch\_RPs \cap DEL\_RPs)$) by the total number of ChEMBL RPs ($N_{total}(Ch\_RPs)$):

$$Ch\_RPs\ cov\ \% = \frac{N_{common}(Ch\_RPs \cap DEL\_RPs)}{N_{total}(Ch\_RPs)}$$

However, the analysis of the percentage of covered ChEMBL RPs does not consider the number of compounds corresponding to each RP, although different RPs can be populated differently – from 1 to ≈12 000 compounds. As a result, increasing RP coverage does not necessarily mean significantly increasing the compound coverage. Thus, the ChEMBL RPs coverage ($W\_Ch\_RPs\ cov\ \%$), weighted by RP population (the number of ChEMBL compounds per RP – $\sum Pop(Ch\_RPs)$ and $\sum Pop(Ch\_RPs \cap DEL\_RPs)$), is also

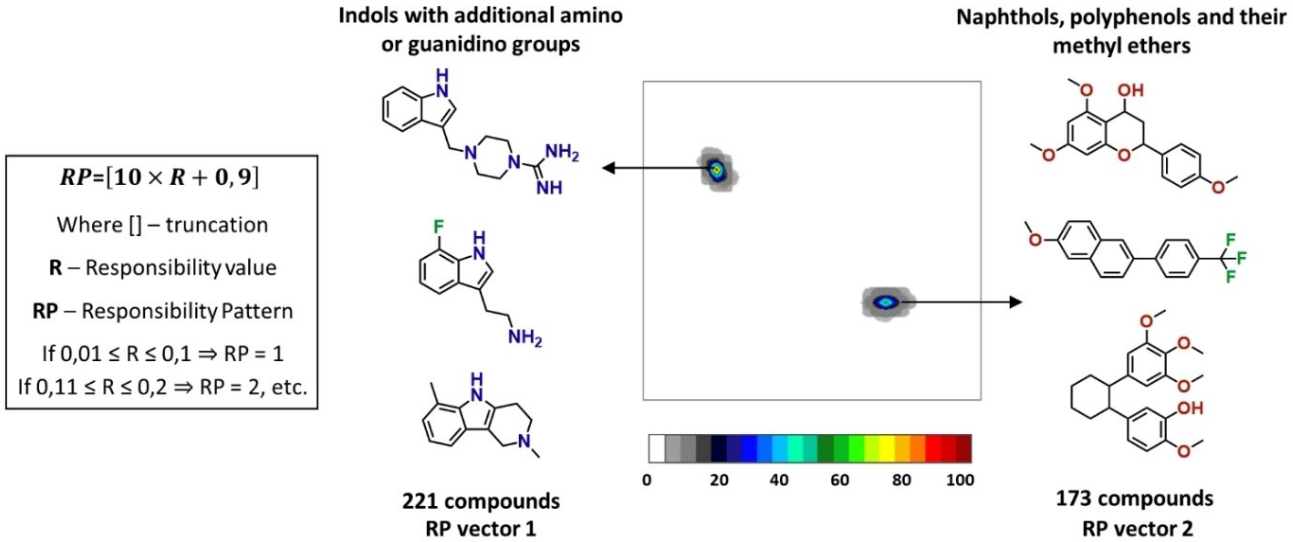


Figure 3. Left: formula for responsibility pattern (RP) calculation. Right: example of compounds sharing the same RPs and their position on the density landscape – a map colored by the local density of compounds. Highly populated zones are colored in red, underpopulated ones – in grey.

used – where *Pop(RP)* stands for the number of ChEMBL compounds featuring that RP:

$$W\_Ch\_RPs\ cov\ \% = \frac{\sum Pop(Ch\_RPs\ \cap\ DEL\_RPs)}{\sum Pop(Ch\_RPs)}$$

3 Data

3.1 Commercially Available BBs

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

3.2 ChEMBL (Biologically Tested Compounds)

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

4 Results and Discussion

4.1 DNA-compatible BBs and Reactions for DEL Generation

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

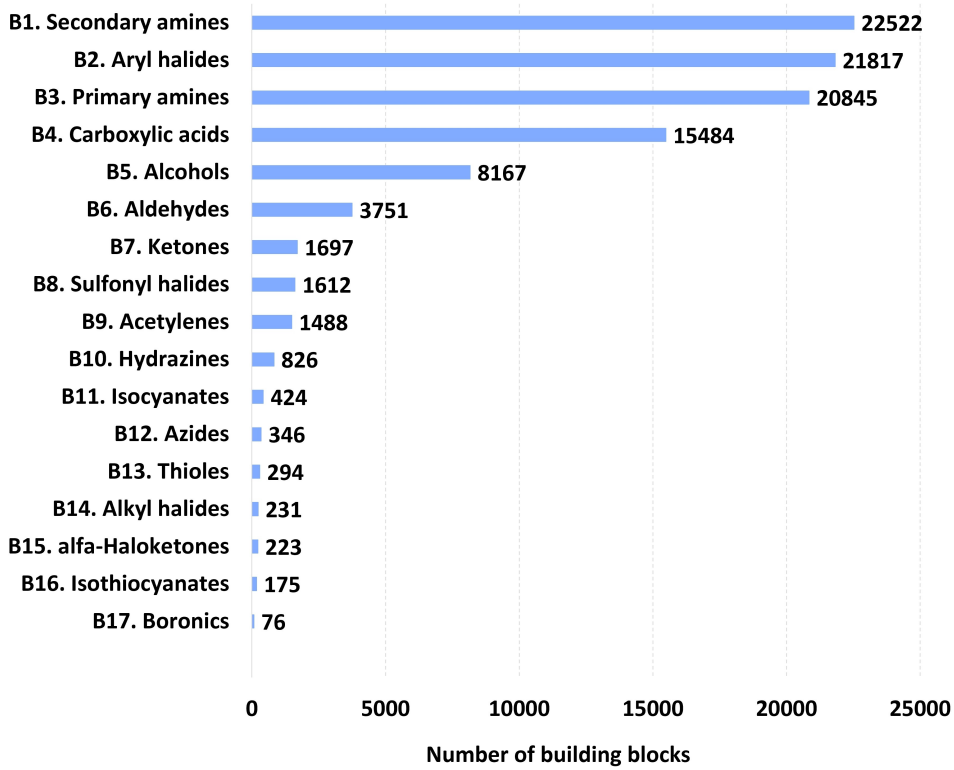


Figure 4. Monofunctional DNA-compatible commercially available BBs.

In this work, 79 141 mono-, bi-, and trifunctional BBs were used for DEL generation. They were obtained by applying the Goldberg rule of two and built-in eDesigner DEL-compatibility filters to the combined in-stock library provided by eMolecules and Enamine. Prevalent monofunctional BB classes in the resulting dataset are secondary and primary amines, aryl halides, and carboxylic acids (Figure 4 and Table S1 in Supporting Information). Due to their participation in common DNA-compatible combinatorial reactions (such as condensation of carboxylic acids with amines, aldehyde reductive amination, bromo-Sonogashira coupling, etc.), there is an active development of such BBs, making these four classes more structurally rich and widely available commercially. Note that in this work, all structures were stereochemistry-depleted (a unique skeleton graph is used to represent all stereoisomers). Therefore, the actual number of different BBs is higher.

In the case of bifunctional BBs (Figure 5 and Table S2 in SI), protected amino acids (AA) (such as amino esters, N-Boc-AA, N-Fmoc-AA, etc.) represent the most abundant class (3 796). The reason for such abundance is the popularity of peptide bond formation for DEL compounds' synthesis that requires this type of reagents. However, the number of actual AA fragments available from BBs with multiple protective groups is slightly smaller (2 885). It appears that the majority of AA fragments (2 173) occur in only one protected form, and 712 AA were found in the library more than once with different protecting groups. Figure 6 (I) shows an example of AAs that occur in the maximum number of protected variations in the BB library.

A similar tendency is also observed for protected diamines that occupy third place in the bar chart in Figure 5 after BBs containing both aryl halide and carboxylic acid functionality (2 359). A total of 737 protected diamines are equivalent to only 632 unique diamine fragments. Among them, 510 are represented by only one protected variant, while the other 122 occur in several differently protected copies. Four diamines, each occurring in the highest number of protected variations, are shown in Figure 6 (II).

The number of trifunctional BBs is significantly lower than other reagents due to higher structural complexity (Figure 7 and Table S3 in SI). The most highly populated class of trifunctional BBs is haloaryl nitrocarboxylic acids containing 110 members. In DEL technology nitro group usually pose as a latent amino group that can be obtained upon reduction.

Using these BBs and user-defined library limitations in eDesigner, 2 497 DELs were designed. The details about these DELs can be found in the Supporting Information (Table S4). The maximal number of heavy atoms in DEL compounds was set to be 45, and at least half of all compounds in the library needed to have less than 35 non-hydrogen atoms. The frequency of the use of a particular reaction to generate all DELs is shown in Figure 8.

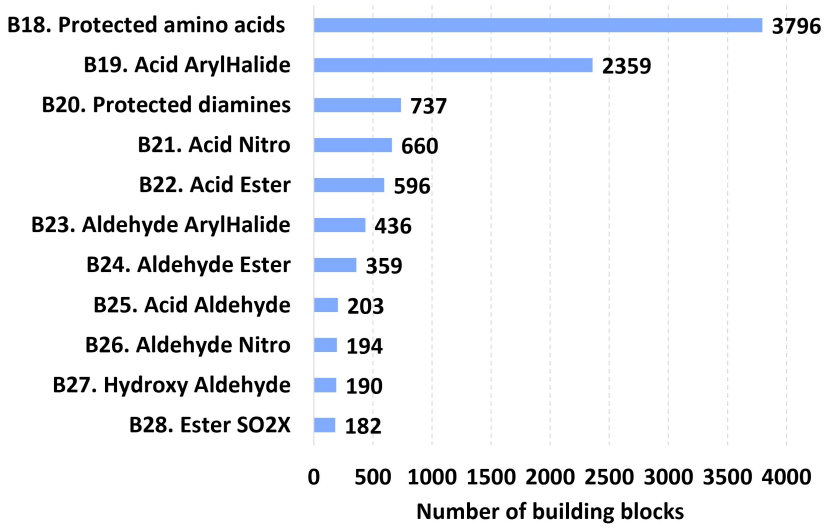


Figure 5. Bifunctional DNA-compatible commercially available BBs.

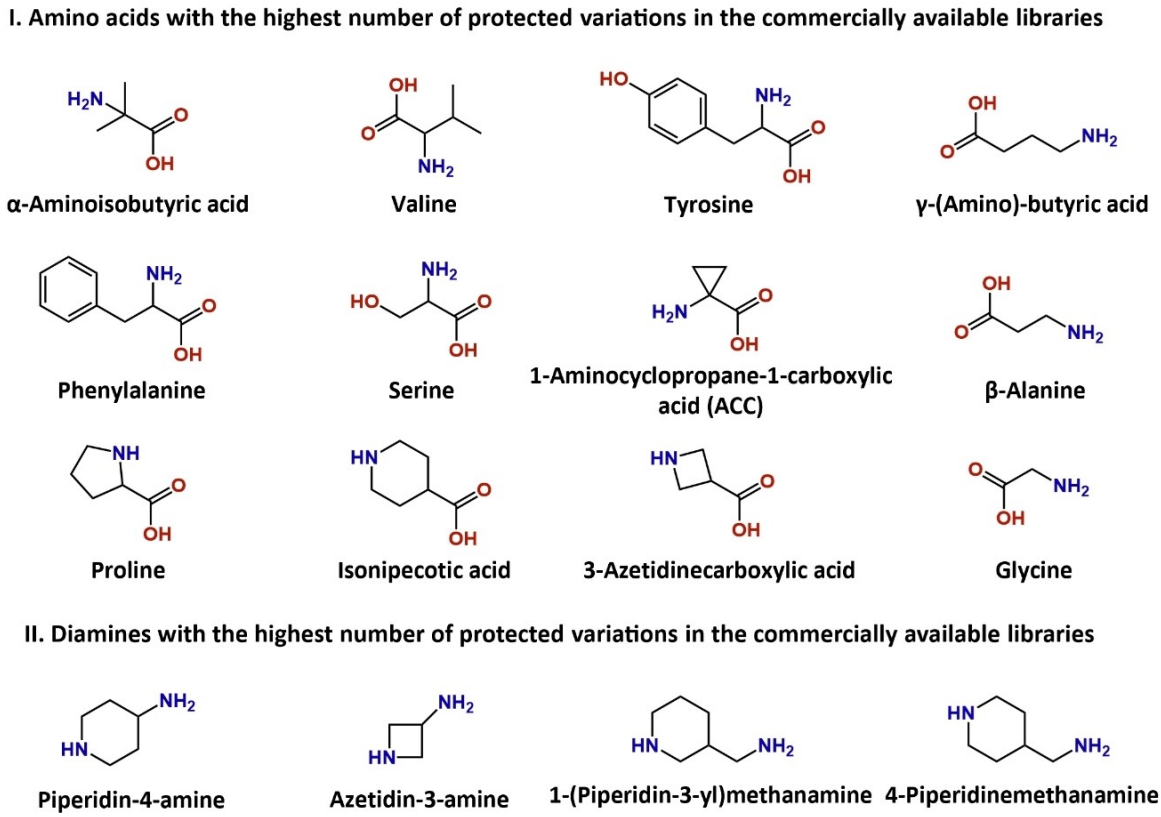


Figure 6. AA (I) and diamines (II), represented in the commercially available libraries of DNA-compatible BBs with the highest number of protected variations (N-Boc, N-Fmoc, various esters, etc.).

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

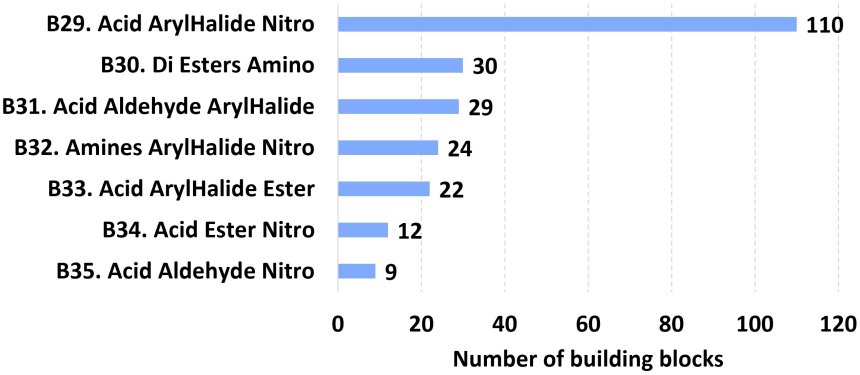


Figure 7. Trifunctional DNA-compatible commercially available BBs.

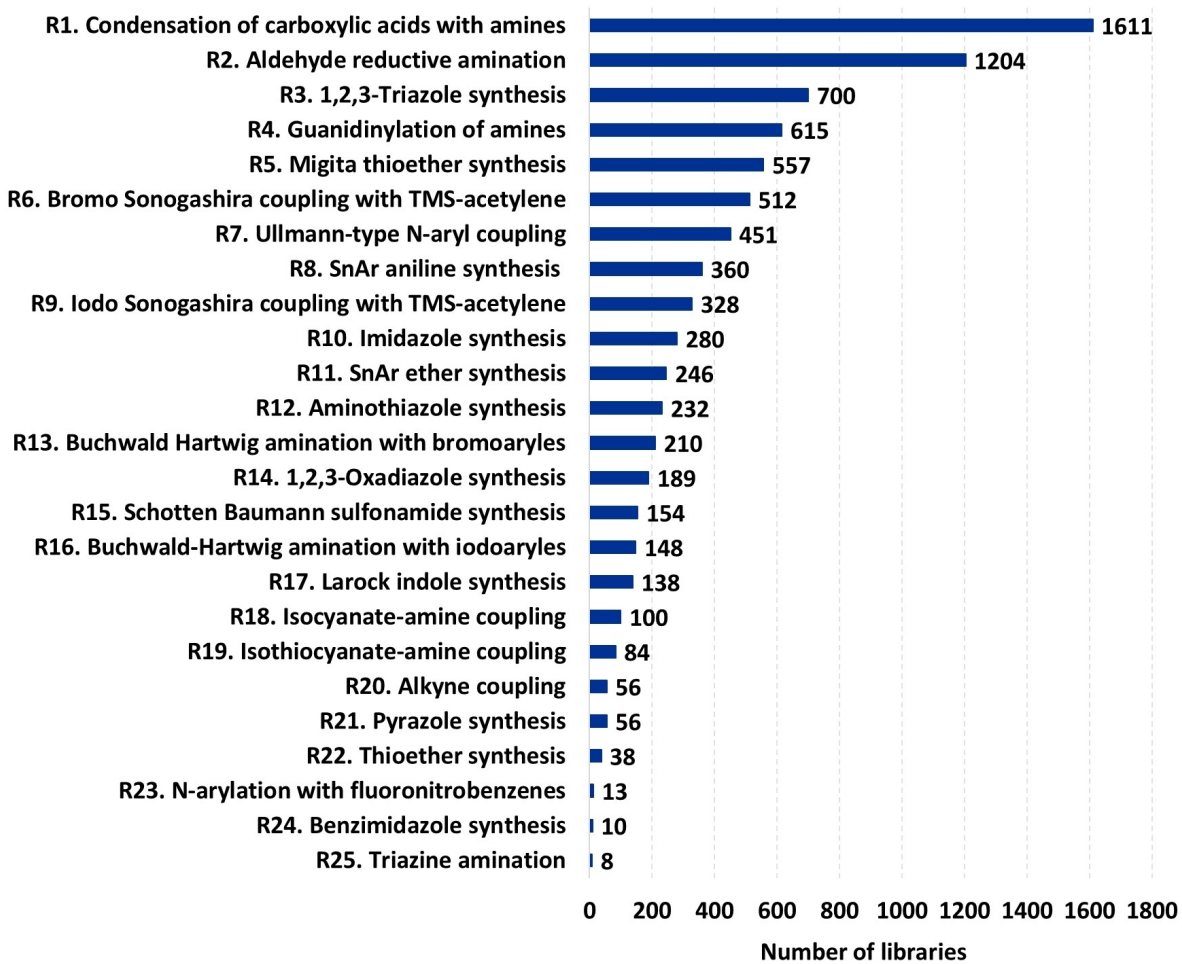


Figure 8. Frequency of the use of a particular reaction in DELs generation.

R4), aryl halides – in two (R5 and R6), and carboxylic acids in R1.

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

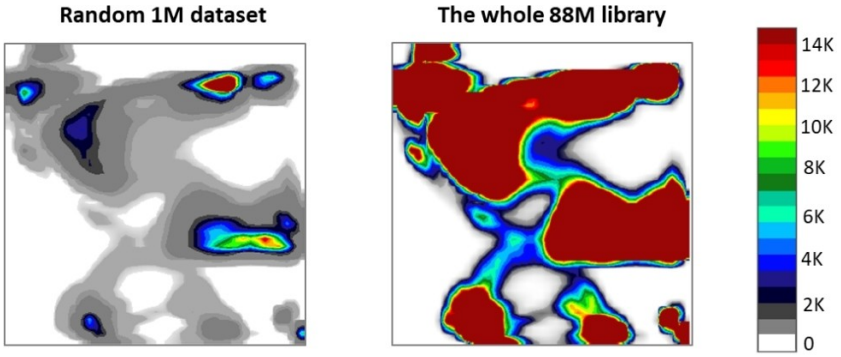


Figure 9. Comparison of the density distribution for the 1 M randomly generated compounds and the whole DEL (88 M). The color scale encodes the corresponding number of compounds residing in each colored node of the map.

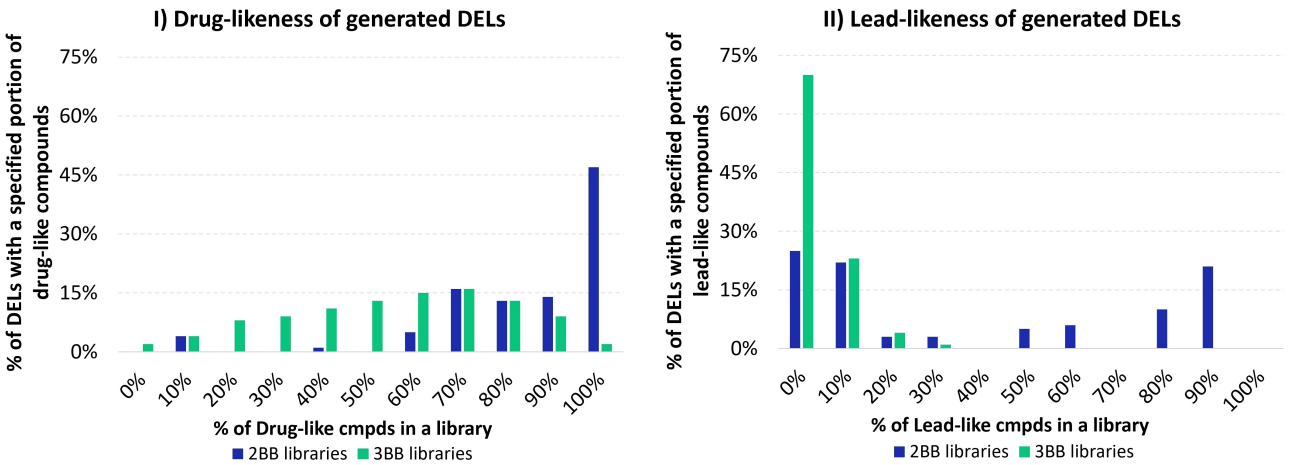


Figure 10. Comparison of (I) drug- and (II) lead-likeness of 2BB and 3BB libraries: percentage of 2BB and 3BB libraries having a particular portion of compounds satisfying respective filters is given.

the density of residents differs. Therefore, 1 M randomly enumerated compounds will be considered in this work as a sufficient representation of the whole DEL for GTM-based analysis.

4.2 Physicochemical Properties of Generated Libraries

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

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

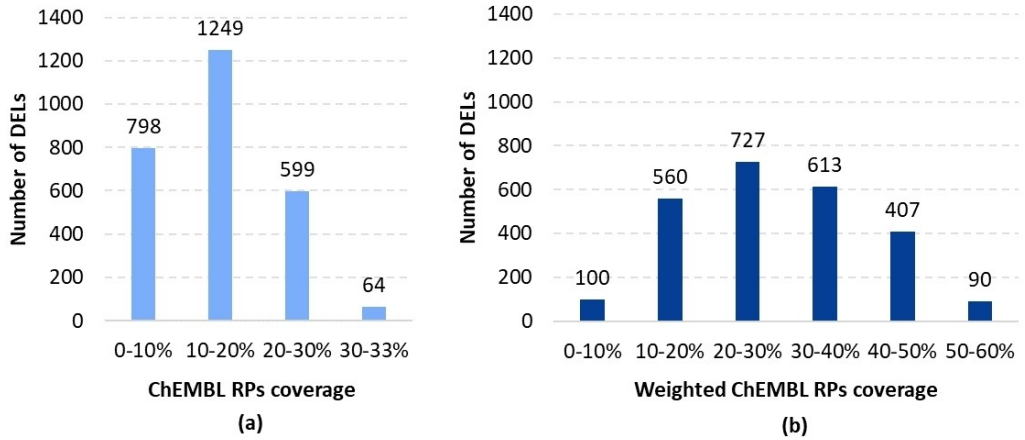


Figure 11. (a) Number of DELs with different coverage of ChEMBL responsibility patterns (RPs) (b) Number of DELs with different percentages of ChEMBL RPs coverage weighted by the RPs population (number of ChEMBL compounds per RP).

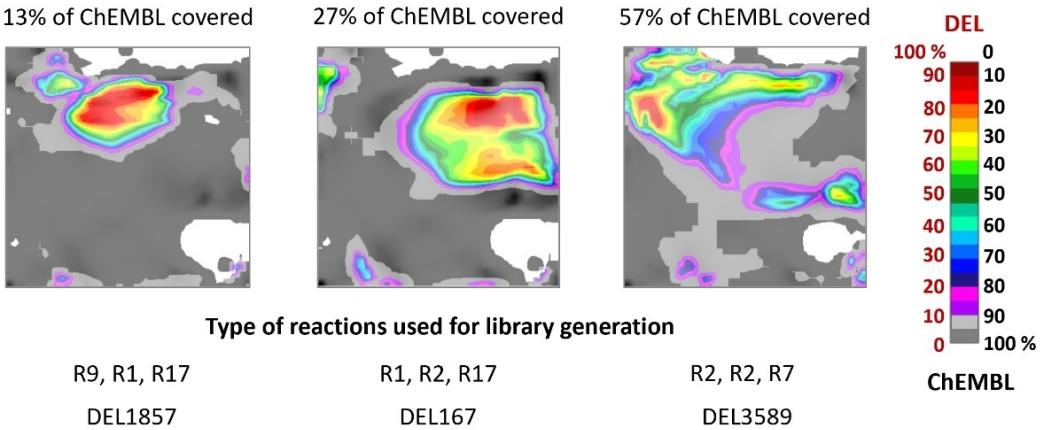


Figure 12. Class landscapes comparing a particular DEL with ChEMBL. From left to right: comparison of ChEMBL to DEL1857, DEL167, and DEL3589. Dark grey zones are populated exclusively by ChEMBL compounds, while all other colors indicate areas also containing DEL compounds in a different ratio. White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

4.3 Search for the "Golden" DEL

The "golden" DEL can be defined as a library that is diverse enough to cover the highest possible proportion of biologically relevant compounds from ChEMBL. This coverage was calculated in terms of common responsibility patterns (RPs) explained in the Methods section. In Figure 11 (a) one can see the number of libraries with particular coverage of ChEMBL RPs. The majority of libraries cover 10–20% of ChEMBL chemical space in terms of unweighted RPs coverage score. 64 DELs showed the highest coverage of ChEMBL RPs – 30–33%. Figure 11 (b) depicts the coverage of the ChEMBL RPs weighted by the number of compounds that correspond to each RP. This time, 90 DELs showed high coverage of ChEMBL chemical space, ranging from 50 to 60%. Five most similar to ChEMBL libraries are shown in Figure S1 in SI.

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

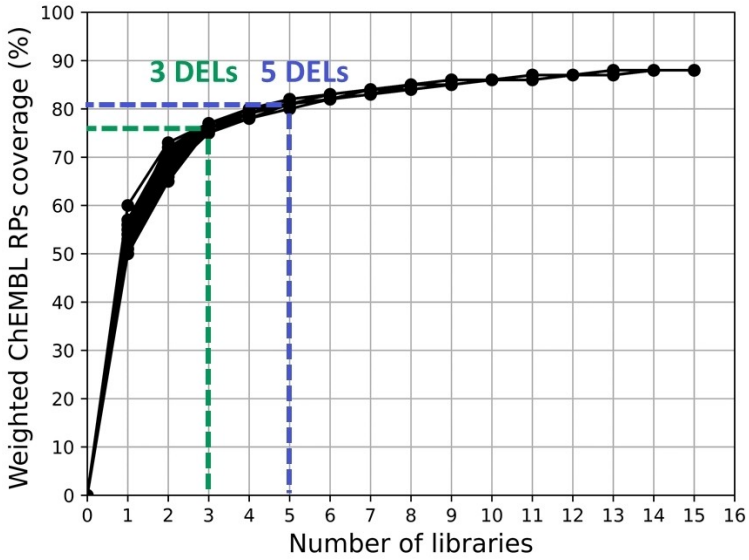


Figure 13. The percentage of the ChEMBL coverage, weighted by the number of compounds sharing common RPs, as a function of the number of libraries in the set. Green and blue dashed lines highlight the points for three and five DELs.

map, but rather distributed on different islands that correspond to different chemotypes, and dark grey areas are less present.

There are 90 libraries with similar chemical space coverage and diversity, but here, we will limit the discussion to the DEL3589 as an example of a "golden" DEL. The 84 M compounds of this DEL can be obtained by a succession of three reactions: two aldehyde reductive amination steps followed by Ullmann-type N-aryl coupling (see Figure S6 in SI). BBs used are 3 138 aldehydes, 275 bromoarylaldehydes, and 97 amines. As was discussed earlier, the latter is the class with the highest number of diverse BBs (Figure 4). Therefore, a random selection of BBs for DEL generation from such various and numerous collection results in higher coverage of ChEMBL chemical space.

4.4 Search for the "Platinum" set of DELs

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

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

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

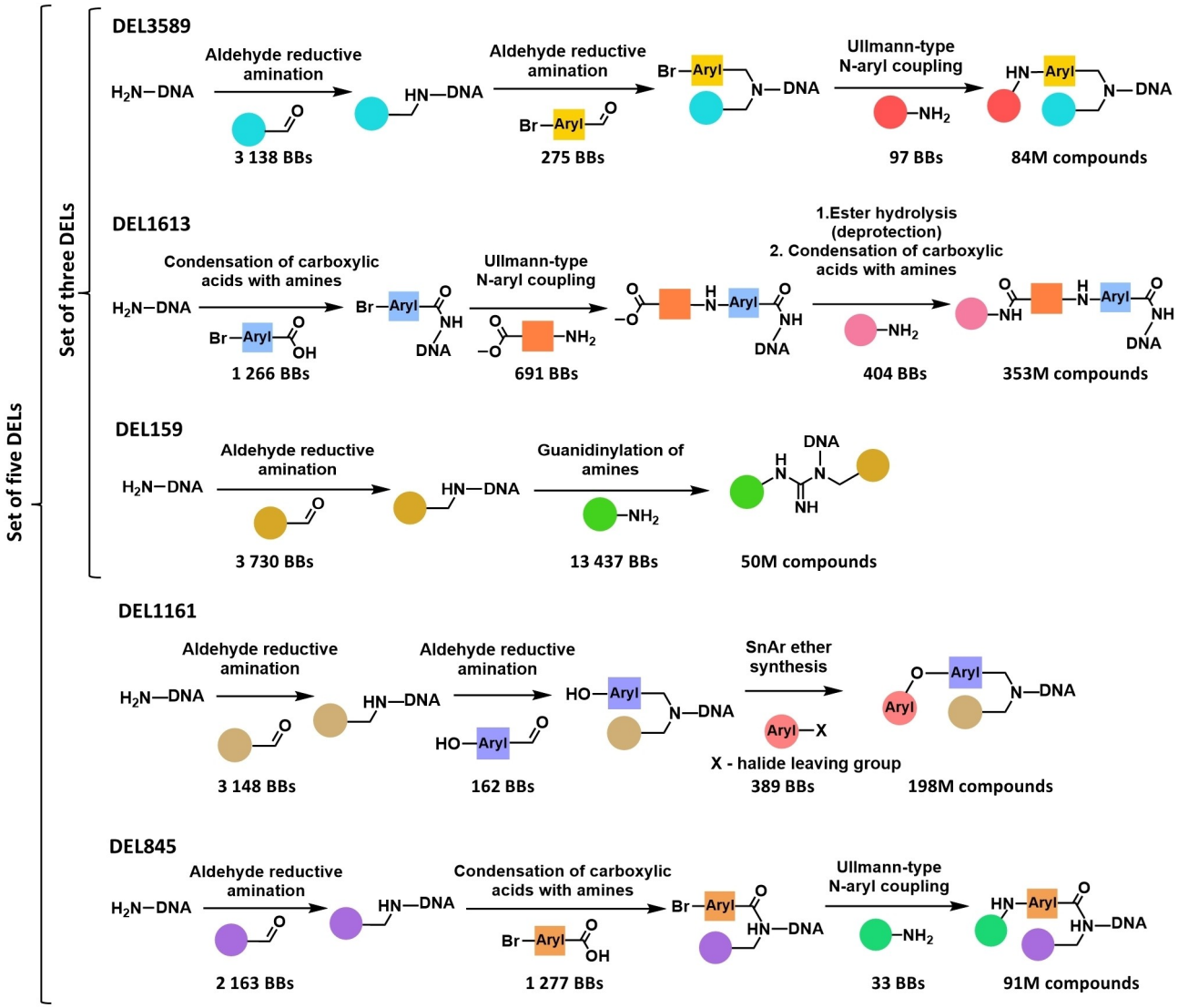


Figure 14. Reactions and BBs required for the synthesis of the "golden" DEL and libraries composing "platinum" pools.

fewer compounds (around 487 M) and yet still allows to cover a large portion of ChEMBL (77%).

The physicochemical properties of the selected libraries were calculated and analyzed (Table 1). The proportion of drug-like and lead-like compounds varies for all DELs. The 2BB DEL159 shows the highest percentage of drug-like and lead-like molecules, 98% and 78%, respectively. This result is not surprising due to the lower molecular weight of compounds from 2BB libraries. Regarding 3BB libraries, it appears that the golden DEL3589 possesses higher drug-likeness (80% of such compounds) and lead-likeness (12% of such compounds) than the 3BB complementary DELs. Indeed, 52% of molecules from DEL1613 are drug-like while for DEL1161 the proportion of such compounds is only 30%. The portion of lead-like molecules for these libraries is negligible. The data on physicochemical properties of compounds from the golden DEL and platinum pools of three and five DELs are available in Figures S2–S4 in SI.

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

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

To better illustrate how ChEMBL coverage increases when a pool of DELs is used instead of a single DEL, four comparative landscapes – featuring the "golden" DEL, the

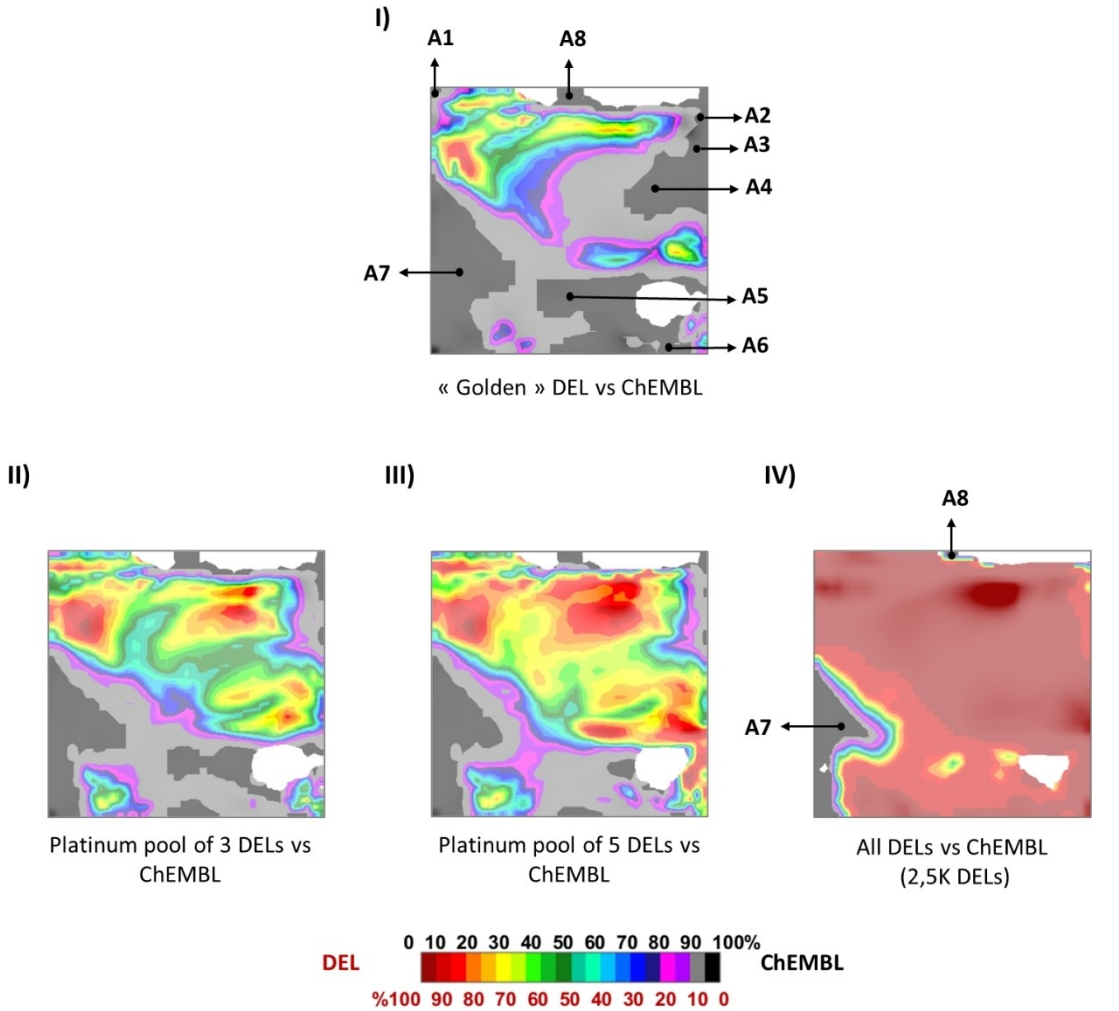


Figure 15. Comparison of ChEMBL and I) "golden" DEL, II) a pool of three DELs, III) a pool of five DELs, and IV) all 2,5 K DELs. Multicolored zones are populated by both ChEMBL and DEL compounds, dark grey zones – only by ChEMBL compounds. White regions correspond to the empty areas of the chemical space.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

"platinum" pools of three and five DELs, and ≈2,5 K DELs against ChEMBL were created (Figure 15). Structural analysis of underrepresented in DELs zones was carried out (Figure 16). The obtained landscapes show that as we go from one (Figure 15 (I)) to three DELs (Figure 15 (II)), the ChEMBL coverage increases drastically. On the landscape of the "platinum" pool of three DELs, the ChEMBL areas from A1 to A6 became a lot more populated. However, the addition of the following two libraries does not have the same impact. There are almost no new previously uncovered areas, only the increase in the population of previously occupied areas is observed (Figure 15 (III)). However, neither three nor five libraries succeeded in covering areas A7 and A8 completely. To see whether it is even possible to do so, a comparative landscape for all DELs versus ChEMBL was created (Figure 15 (IV)). It appears that neither of the DELs can cover these regions of the chemical space – areas A7 and A8 remained dark-grey. This result is not surprising because they contain natural products (NP) and NP-like compounds such as cardiac glycosides, steroids, and steroid-like compounds, saccharides, nucleotides, oligopeptides, coumarins, macrolides, chalcones, etc., which are indeed inaccessible by DEL technology as employed in this analysis.

5 Conclusions

In this work, for the first time, the ultra-large chemical space of DNA-encoded libraries (DELs) containing 2,5B compounds in total (2.5 K libraries 1 M each) was designed and generated using eDesigner and analyzed with the help of GTM. Owing to the probabilistic nature of GTM and efficiency of the libraries analysis and comparison based on the responsibility patterns, it was possible to develop a GTM-based approach for quick selection of DELs occupying

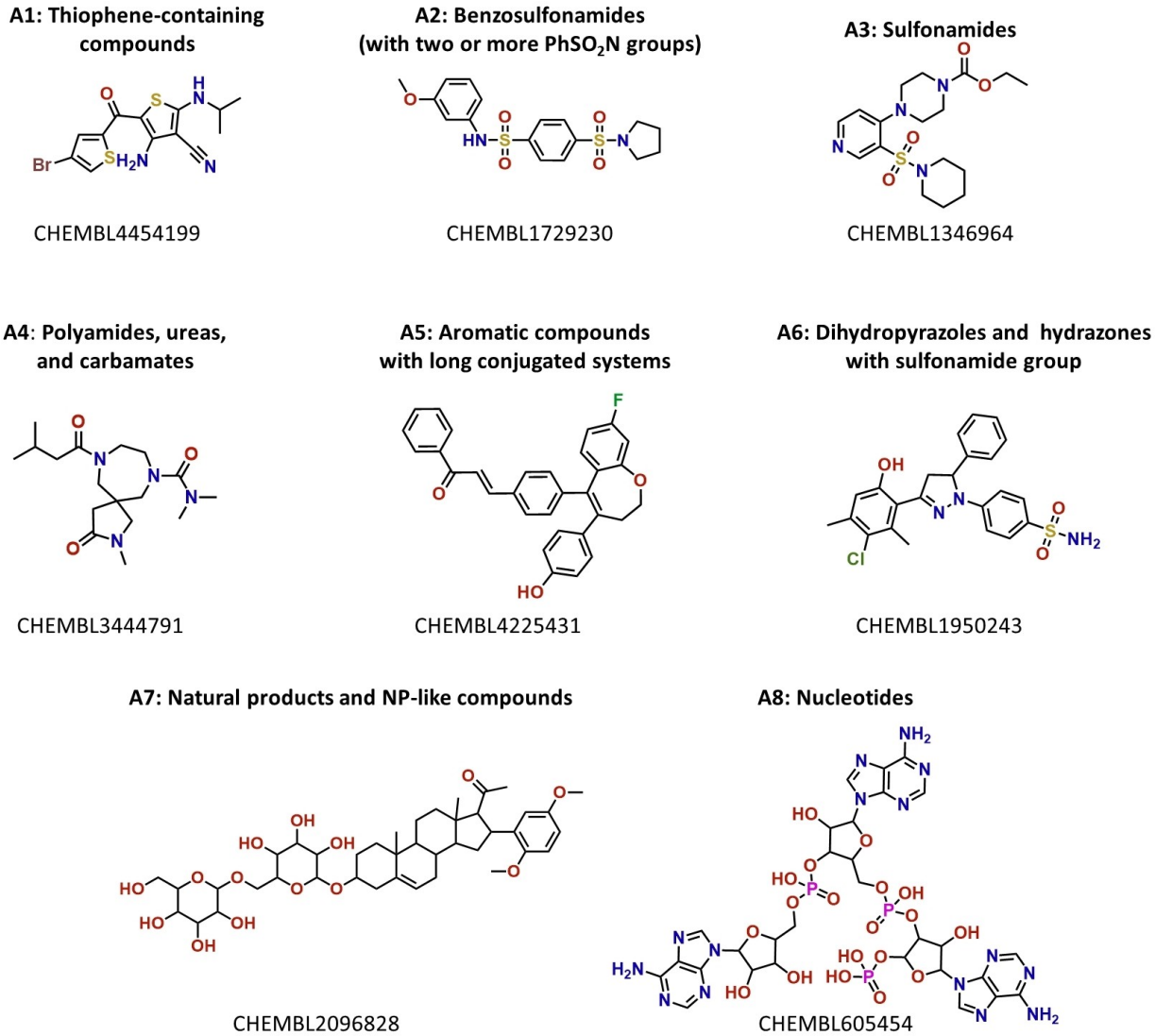


Figure 16. Examples of ChEMBL compounds populating areas from A1 to A8 highlighted in landscapes in Figure 15.

the same areas of the chemical space as the reference library. In this work, the goal was to detect the "golden" DEL or "platinum" pool of DELs for primary screening – the libraries containing the highest portion of biologically relevant chemotypes. Therefore, ChEMBL, as the largest database of dose-response activity tests and thus an optimal representation of biologically relevant space, was used as a reference. However, the approach described herein could be applied to any reference library, e.g., actives of a particular biological target.

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

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

Author Contribution

The manuscript was written through the contribution of all authors and all of them approved the final version of the manuscript.

Acknowledgments

The authors are grateful to eMolecules Inc.[24] for the provided library of commercially available BBs, used for DNA-encoded libraries design.

Conflict of interest

None declared.

Data Availability Statement

The data used in this work are available in the public domain resources: biologically relevant compounds from ChEMBL[13] (version 28) – https://www.ebi.ac.uk/chembl/, collection of eMolecules[24] building blocks partially available on the site https://www.emolecules.com/info/products-building-blocks.html, collection of Enamine[25] building blocks available on the site https://enamine.net/building-blocks.

SMILES file for the golden DEL is available in Supporting Information.

References

[1] a) M. S. Attene-Ramos, C. P. Austin, M. Xia, in *Encyclopedia of Toxicology*, **2014**; b) J. Inglese, D. S. Auld, in *Wiley Encyclopedia of Chemical Biology*, **2008**.
[2] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, G. S. Sittampalam, *Nat. Rev. Drug Discovery* **2011**, *10*, 188–195.
[3] R. M. Franzini, C. Randolph, *J. Med. Chem.* **2016**, *59*, 6629–6644.
[4] N. Favalli, G. Bassi, J. Scheuermann, D. Neri, *FEBS Lett.* **2018**, *592*, 2168–2180.
[5] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, A. Chuprina, K. E. Gubina, Y. S. Moroz, *iScience* **2020**, *23*, 101681.
[6] S. Brenner, R. A. Lerner, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 5381–5383.
[7] R. A. Goodnow Jr, in *A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery*, John Wiley & Sons, **2014**.
[8] A. L. Satz, *ACS Med. Chem. Lett.* **2018**, *9*, 408–410.
[9] a) R. M. Franzini, D. Neri, J. Scheuermann, *Acc. Chem. Res.* **2014**, *47*, 1247–1255; b) D. Madsen, C. Azevedo, I. Micco, L. K. Petersen, N. J. V. Hansen, *Prog. Med. Chem.* **2020**, *59*, 181–249.
[10] D. T. Flood, C. Kingston, J. C. Vantourout, P. E. Dawson, P. S. Baran, *Isr. J. Chem.* **2020**, *60*, 268–280.
[11] A. Kontijevskis, *J. Chem. Inf. Model.* **2017**, *57*, 680–699.
[12] A. Martin, C. A. Nicolaou, M. A. Toledo, *Commun. Chem.* **2020**, *3*, 1–9, ARTN 127.
[13] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Felix, M. P. Magarinos, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Maranon, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic Acids Res.* **2019**, *47*, D930–D940.
[14] C. M. Bishop, M. Svensen, C. K. I. Williams, *Neural Computation* **1998**, *10*, 215–234.
[15] Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, A. Varnek, *J. Chem. Inf. Model.* **2021**, *61*, 179–188.
[16] I. Casciuc, Y. Zabolotna, D. Horvath, G. Marcou, J. R. Bajorath, A. Varnek, *J. Chem. Inf. Model.* **2018**, *59*, 564–572.
[17] F. W. Goldberg, J. G. Kettle, T. Kogej, M. W. D. Perry, N. P. Tomkinson, *Drug Discovery Today* **2015**, *20*, 11–17.
[18] a) F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855–868; b) A. Varnek, D. Fourches, V. Solov'ev, V. E. Baulin, A. N. Turanov, V. K. Karandashev, D. Fara, A. R. Katritzky, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1365–1382.
[19] K. Klimenko, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2016**, *56*, 1438–1454.
[20] Y. Zabolotna, D. M. Volochnyuk, S. V. Ryabukhin, K. Gavrylenko, D. Horvath, O. Klimchuk, O. Oksiuta, G. Marcou, A. Varnek, *J. Chem. Inf. Model.* **2021**, DOI: 10.1021/acs.jcim.1c00754.
[21] *LillyMol: Eli Lilly Computational Chemistry and Chemoinformatics Group Toolkit*, **2020**, https://github.com/EliLillyCo/LillyMol.
[22] D. Horvath, G. Marcou, A. Varnek, *Drug. Discov. Today Technol.* **2019**, *32–33*, 99–107.
[23] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, *J. Comput.-Aided Mol. Des.* **2015**, *29*, 1087–1108.
[24] eMolecules Inc., https://www.emolecules.com/.
[25] Enamine Ltd., https://enamine.net/.
[26] *ChemAxon. JChem*, **2020**, Version 20.8.3, ChemAxon Ltd: Budapest, Hungary.
[27] *Virtual Screening Web Server 2020*, http://infochim.u-strasbg.fr/webserv/VSEngine.html.
[28] C. Zambaldo, S. N. Geigle, A. L. Satz, *Org. Lett.* **2019**, *21*, 9353–9357.
[29] A. L. Satz, J. Cai, Y. Chen, R. Goodnow, F. Gruber, A. Kowalczyk, A. Petersen, G. Naderi-Oboodi, L. Orzechowski, Q. Strebel, *Bioconjugate Chem.* **2015**, *26*, 1623–1632.
[30] G. Landrum, *RDKit: Open-Source Cheminformatics Software*, **2016**, http://www.rdkit.org/, https://github.com/rdkit/rdkit.
[31] C. A. Lipinski, *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235–249.
[32] M. P. Gleeson, *J. Med. Chem.* **2008**, *51*, 817–834.
[33] a) O. Eidam, A. L. Satz, *MedChemComm* **2016**, *7*, 1323–1331; b) Z. Wu, T. L. Graybill, X. Zeng, M. Platchek, J. Zhang, V. Q. Bodmer, D. D. Wisnoski, J. Deng, F. T. Coppo, G. Yao, *ACS Comb. Sci.* **2015**, *17*, 722–731.

Received: October 23, 2021
Accepted: January 3, 2022
Published online on January 28, 2022

Correction added on July 21 2022 after first online publication: Updation in the license signed by the author.

Summary

In this work, 2.5K virtual DELs were generated, and their chemical spaces were analyzed to identify an optimal library for primary screening. The GTM-based approach for quick comparison of libraries to a reference ChEMBL28 database was developed. As an optimal representation of the reference library for primary screening, the ChEMBL28 database, the largest repository of dose-response activity assays, was selected for its diversity and biological relevance. The proposed GTM-based coverage scores allowed to measure the chemotype similarity between DELs and ChEMBL, speeding up the process of library comparison. This approach allowed to identify a DEL covering 57% of ChEMBL chemical space. However, such coverage is still quite low. Consequently, the strategy was to select a set of complementary DELs to cover different regions of ChEMBL, thus achieving higher overall coverage. DELs to include in the optimal set were identified iteratively – each DEL was selected in a way so that it covers ChEMBL zones that were not covered in the previous steps. In this way, sets of five and three DELs were selected that together provide the highest coverage of ChEMBL chemical space, covering it by 81% and 77%, respectively.

The analysis of the comparative landscape of ChEMBL vs 2.5K DELs allowed to conclude that even all DELs together cannot cover 100% of ChEMBL space. The uncovered areas were investigated and found to contain natural products (NPs) or NP-like compounds, such as steroids, macrolides, and nucleotides. Such compounds are expectedly not covered by herein-generated DELs, since the design of NP-like DELs was not the goal of the study. In addition, drug- and lead-likeness of generated DELs were estimated according to the theoretically calculated physicochemical properties of their compounds. As anticipated, 2BB libraries contain smaller compounds, resulting in a higher proportion of drug-like and lead-like molecules compared to 3BB DELs – only 3% of them are fully drug-like.

This work represents a pioneering study on the exploration and structural analysis of the space of DELs at such a scale. Findings from this study and the space itself can be useful for medicinal chemists working with DEL chemistry.

5. Chemical Library Space (CLS) analysis

Introduction

DELs introduced new challenges for chemical library design and analysis. Due to their combinatorial nature, thousands of various DEL designs can be created but not any of them is worth synthesis and biological testing. Instead, possible DELs should be virtually generated and compared to identify the one that justifies the investment of resources for its screening. In addition, DELs are not cherry-pickable – once a library is synthesized, it is stored as a mixture in an Eppendorf tube. Therefore, DELs should be considered and handled in their entirety, rather than being broken down into individual parts. They can also be ultra large-sized – containing up to trillions of compounds[24], thus necessitating 'big data' compatible approaches for their analysis. All these factors influence how DELs should be treated chemoinformatically to select the most promising library for a drug discovery task.

Glossary

CRV – Cumulated Responsibility Vector that is derived from the GTM of the library. It is a vector that encodes the approximate total number of compounds from each node of the map. It is used as a representation of the library as a whole and can be called a CLS vector.

Normalized CRV (Φ) – Library-size independent library descriptor vector. It can also be visualized as a density landscape showing the quantitative distribution of compounds in the chemical space of a library.

Library-modulated CRV (Λ) – A vector representing a library with respect to its overlap with a reference collection. It can be visualized as a library comparative landscape.

Property-modulated CRV (Ω) – A vector allowing to represent both the chemotype and property distribution inside the library. It can be visualized as a property landscape.

In this work, a vectorial-based representation of chemical libraries was proposed to enable more efficient and rapid comparison of compound collections such as DELs. The concept of Chemical Library Space (CLS) was introduced to represent a space that encompasses entire compound libraries as residing objects. This space can be encoded using various library descriptors, with library descriptor vectors defining the position of compound collections within it.

A library descriptor vector can be derived from a GTM-generated map of a compound collection. GTM is a probabilistic dimensionality reduction method defining the position of a data point (molecule) on the latent space map by probabilities specific to map nodes called "responsibilities". The sum of responsibility vectors over all compounds from the library provides a Cumulative Responsibility Vector (CRV) that allows to encode the entire chemical space of a library. Different variations of this vector were proposed here to encode compound collections from many perspectives. To encode the structural composition of the library the Normalized CRV (Φ) was created. To encode the structural composition with respect to some reference library a Library-modulated CRV (Λ) was proposed. Finally, a property-modulated CRV (Ω) allowing to encode a library both in terms of its structural and property distribution was created. Two additional fingerprint and vectorial representations (Γ and Γw), based on responsibility patterns[71] (RPs) proposed in our previous study[10], were used as benchmarks. RP-based representation enabled the encoding of the chemotype composition of the library.

The introduced library vectorial representations (Φ, Λ, and Ω) were used herein to rank DELs by their similarity to ChEMBL28 both in terms of structural composition and property distribution. To do this, a Tanimoto similarity coefficient between each DEL-ChEMBL vector pair was calculated.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

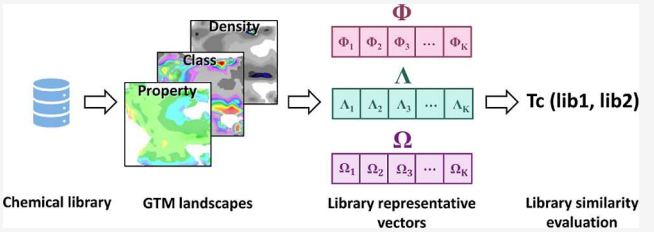
Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case

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

Cite This: J. Chem. Inf. Model. 2023, 63, 4042−4055 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 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.

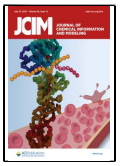


1. INTRODUCTION

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 ultralarge 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

Received: April 4, 2023
Published: June 27, 2023



https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

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 enabling the comparison of 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 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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

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 in a detailed investigation of its compound space where each compound is defined by molecular descriptors—in our case ISIDA fragment counts.[18] 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 responsibility (density) for each of 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 compounds may not have exactly the same responsibility vector. However, similar compounds of a same "chemotype",

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

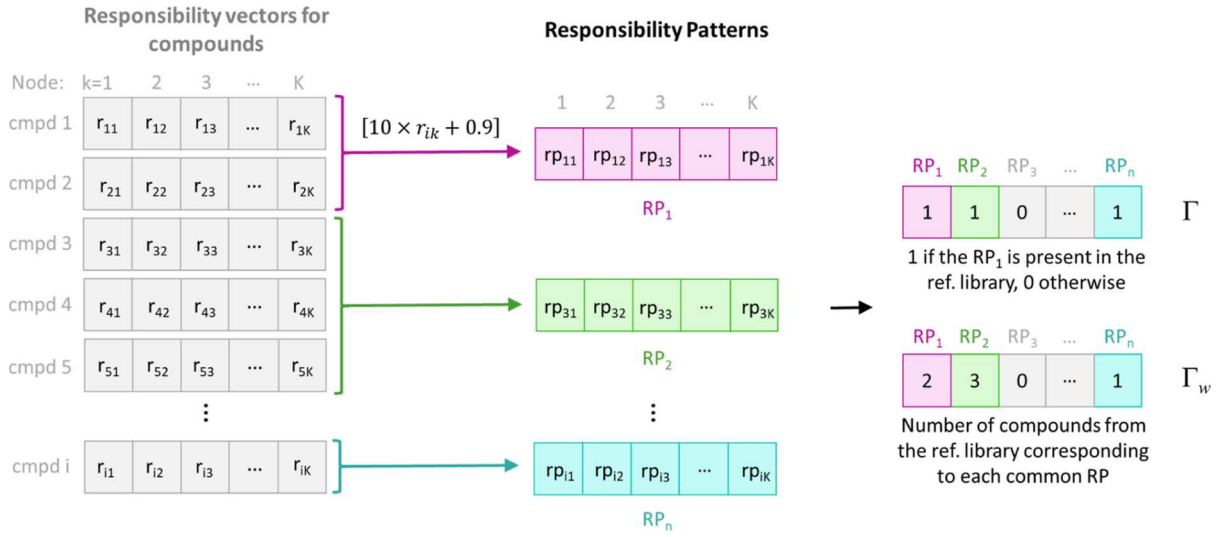


Figure 1. Summary of the RP-based library representations. The Γ values for a particular library are assigned based on the presence or absence of a certain RP in the reference library, and the Γ_w values represent the counts of reference library compounds covered by this RP.

mapping not with identical, but roughly the same rp_{ik} values may be clustered together under a same Responsibility Pattern, defined in eq 1. Therefore, RP[20] distributions within a library implicitly reflects the chemotype distribution, without the need to explicitly predefine "chemotypes".

$$rp_{ik} = [10 \times r_{ik} + 0.9] \tag{1}$$

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.

$$c_k = \sum_{i}^{N} r_{ik} \tag{2}$$

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.

$$\Phi_k = \frac{c_k}{N} \tag{3}$$

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 $1 \le \Lambda_k \le 2$ 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 $1 < \Lambda_k < 2$.

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

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.

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

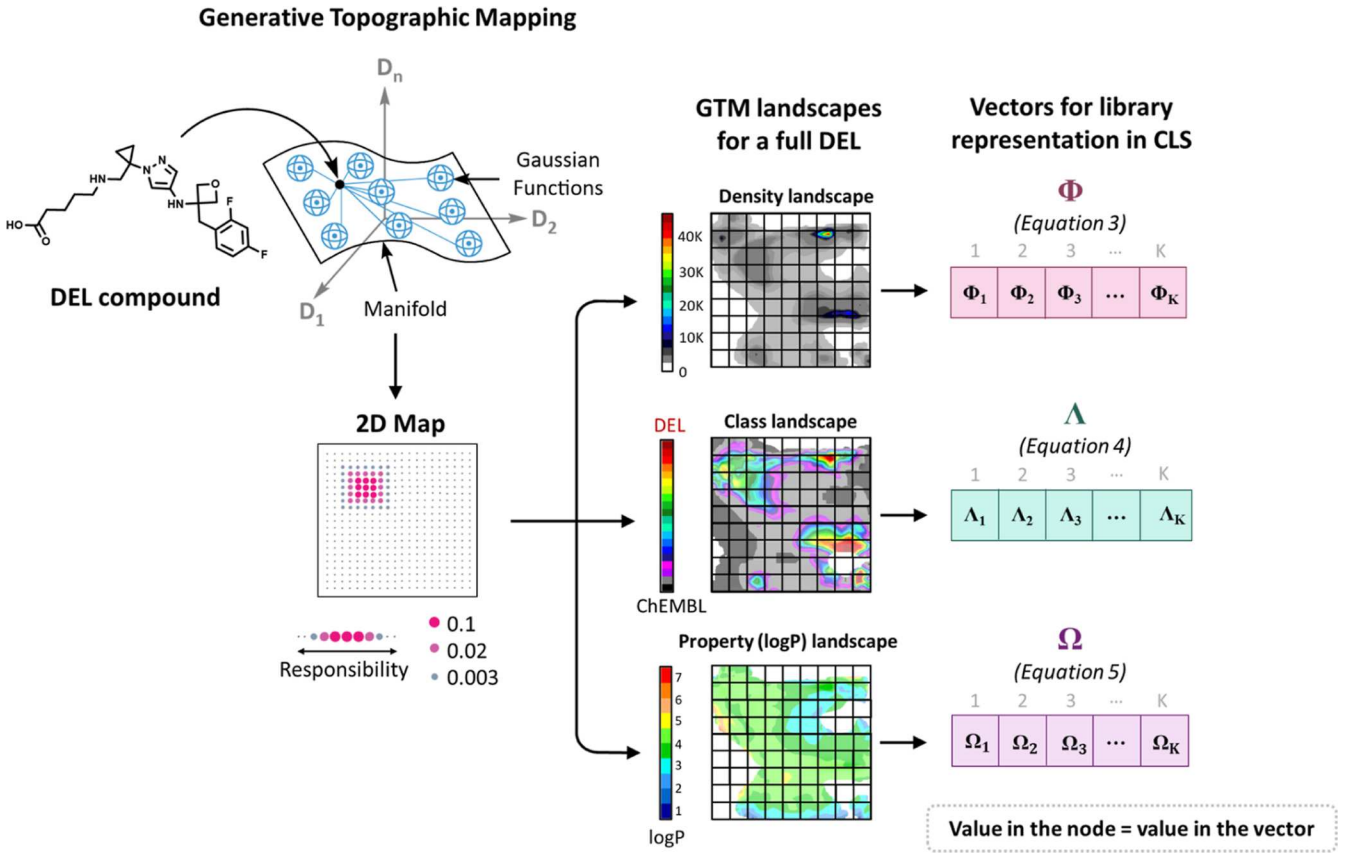


Figure 2. Scheme depicting how each of the introduced herein library encodings (Φ, Λ, and Ω) are derived from the GTM for a particular compound library.

When aiming to maximize representativity 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.

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

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.

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

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.

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

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.

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

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 their 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 representativity of the subset.

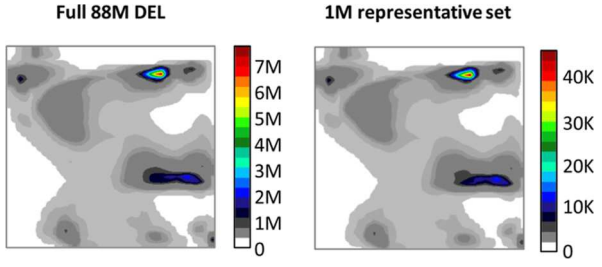


Figure 3. Density landscapes of the entire 88M DEL2568 and its 1M representative subset.

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

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

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 in 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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

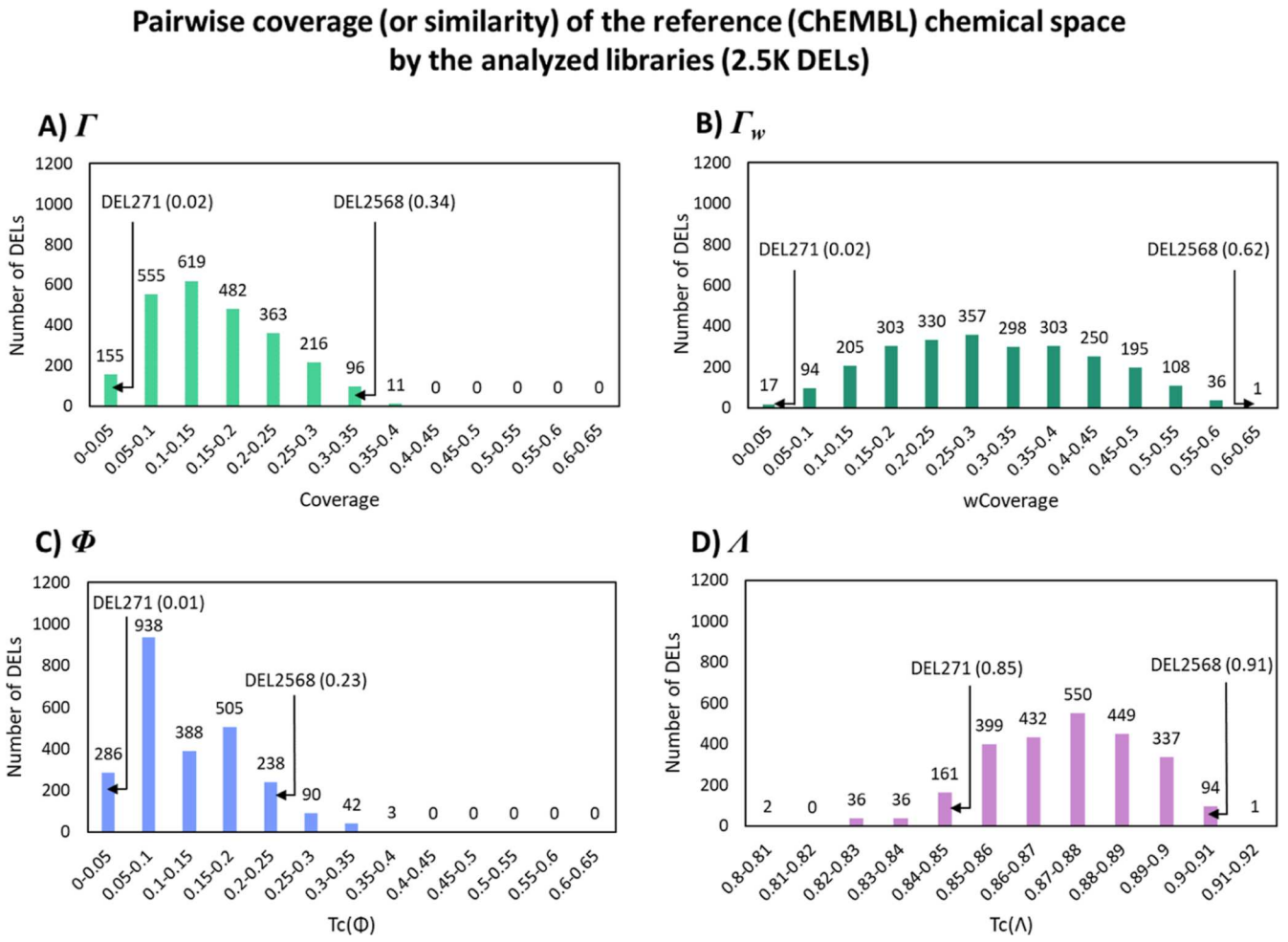


Figure 4. Pairwise comparison of 2.5k DELs with ChEMBL using different representations and metrics: distribution of ChEMBL coverage scores calculated using Γ (A) and Γ_w (B), and distribution of Tc between ChEMBL and each DEL calculated using Φ (C) and Λ (D).

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, 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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

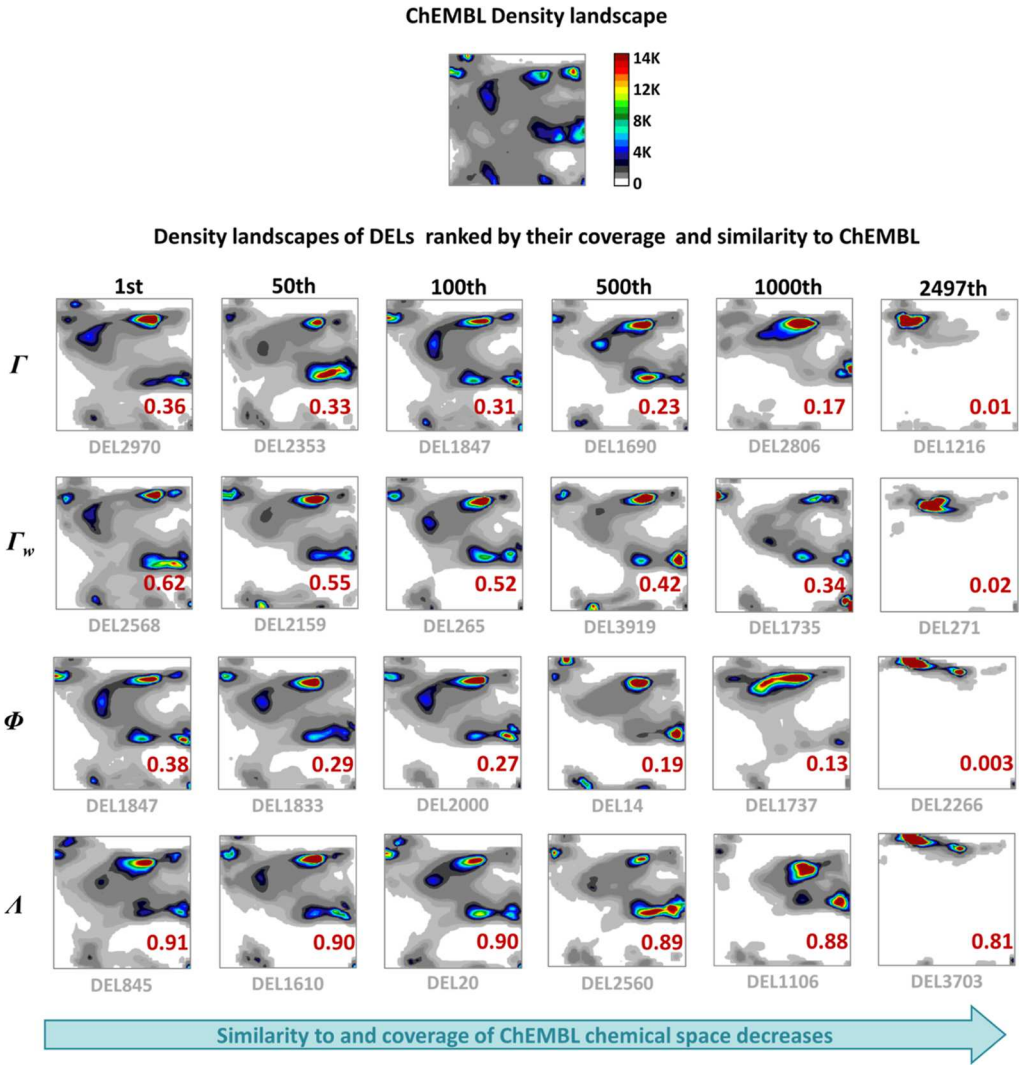


Figure 5. Density GTM landscapes of ChEMBL28 and selected DELs ranging from the most similar to the least similar to ChEMBL. DELs were selected and ranked either by coverage scores (in the case of Γ and Γ_w) or Tanimoto similarity coefficients (in the case of Φ and Λ). Values of either coverage or Tc 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, visual inspection shows that the first few hundred DELs have similar density landscapes to those of the top-ranked library—landscapes corresponding to the 100th or even 500th-ranked library still match landscape No. 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,

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

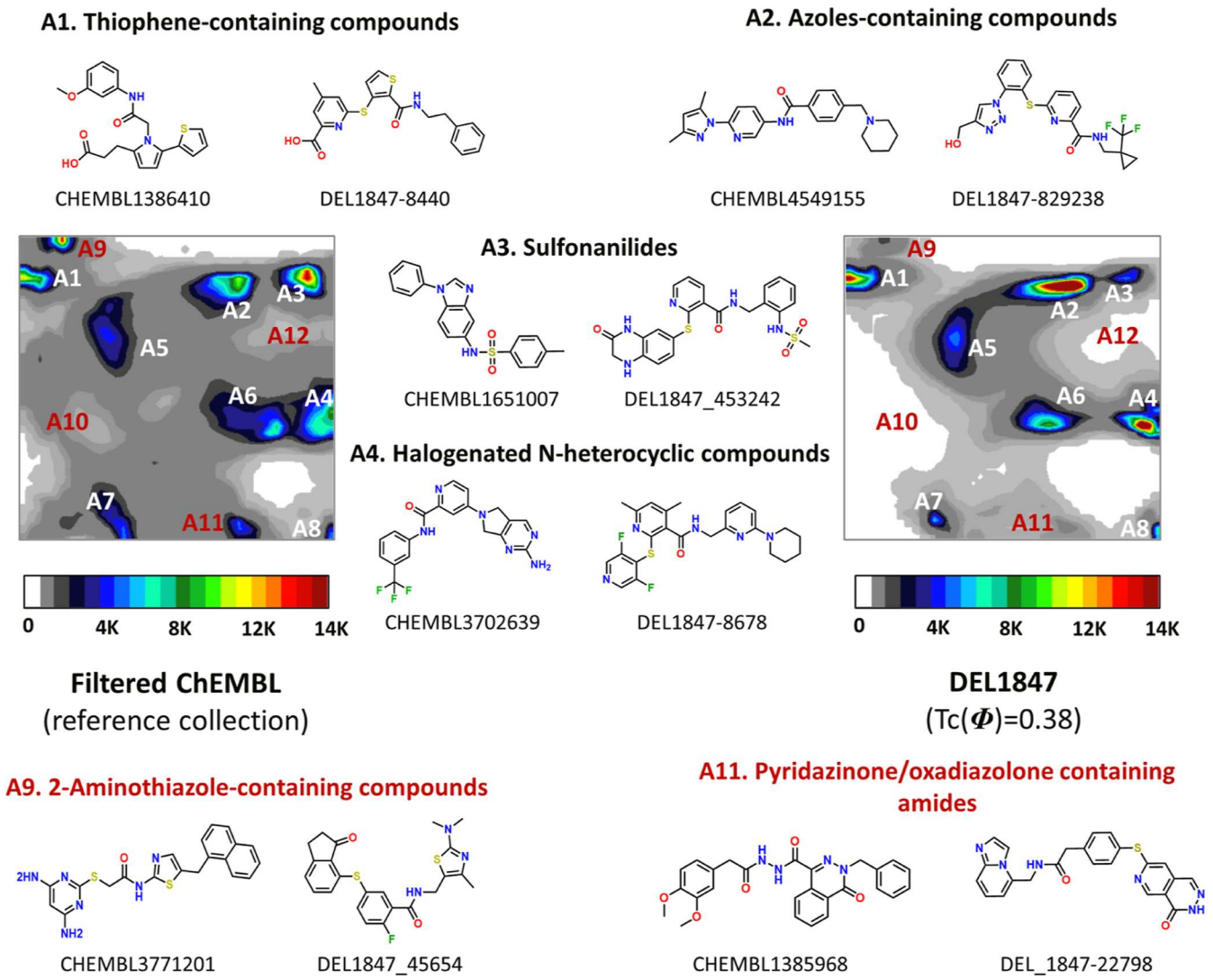


Figure 6. Interpretation of the similarity between ChEMBL and DEL1847 *via* structural analysis of the density landscapes of those libraries. Areas A1−A8 (labeled in white) correspond to the peaks of high density in ChEMBL space that were reproduced in DEL1847. Areas A9−A12 (labeled in red) represent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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

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 always 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.

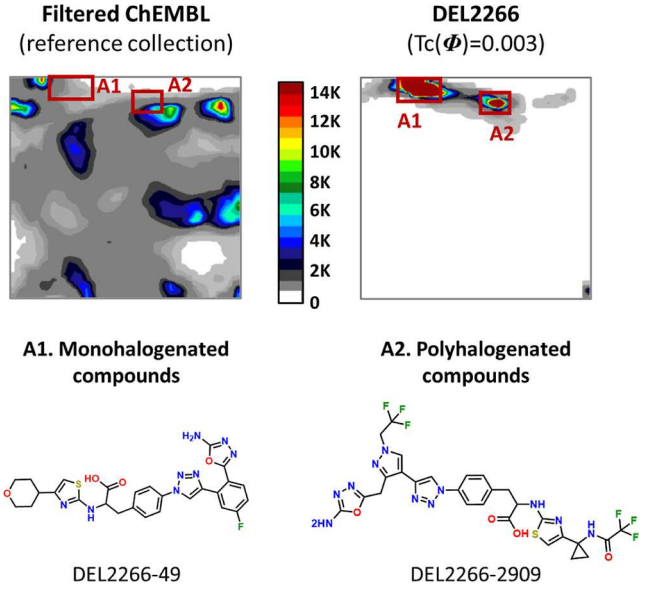


Figure 7. Interpretation of the similarity between ChEMBL and DEL2266 *via* structural analysis of the density landscapes of these libraries.

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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

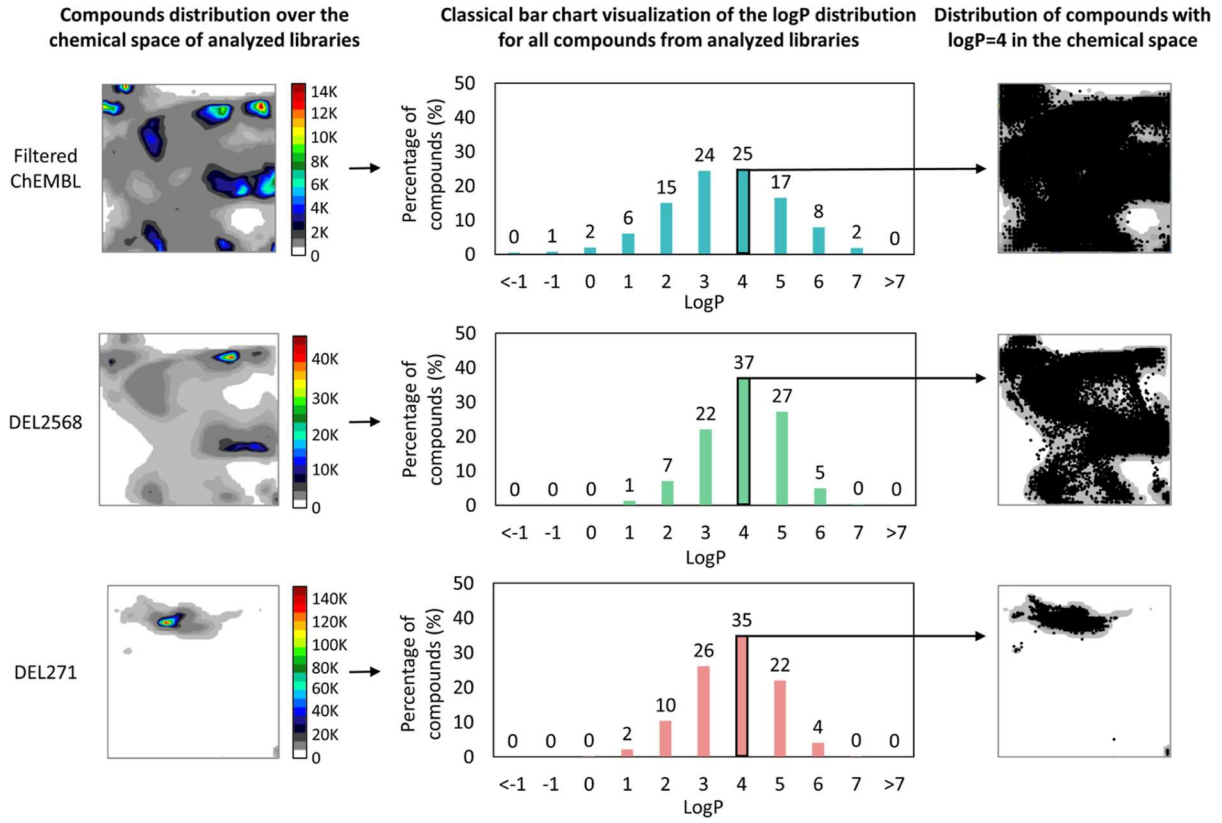


Figure 8. (Left) Density landscapes of filtered ChEMBL, DEL2568, and DEL271; (center) classical bar chart visualization of calculated log *P* distribution for all compounds from analyzed libraries; (right) compounds with log *P* = 4 (black dots) projected on the corresponding density landscapes.

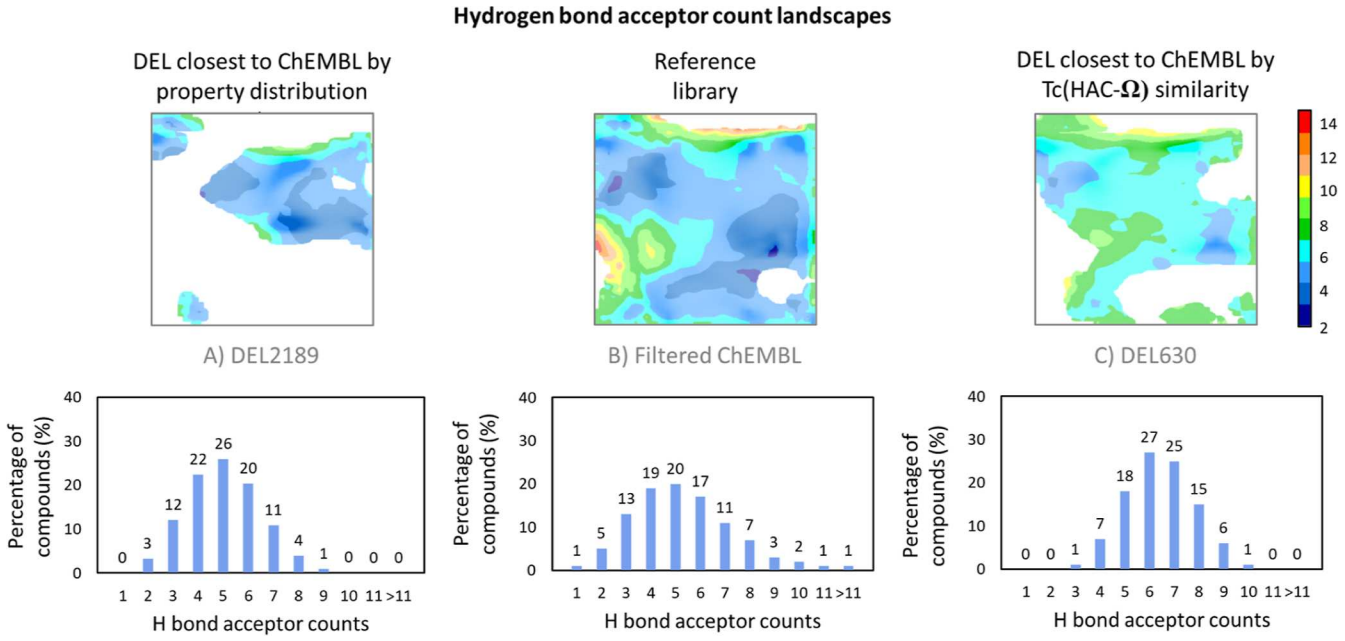


Figure 9. Hydrogen bond acceptor count (HAC) landscapes for (A) DEL2189 (selected by property distribution similarity), (B) reference library-filtered ChEMBL28, and (C) DEL630 [selected by Tc(HAC-Ω) similarity].

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 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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

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 DNA-encoded libraries (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 Chemical Library Space (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 pattern 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

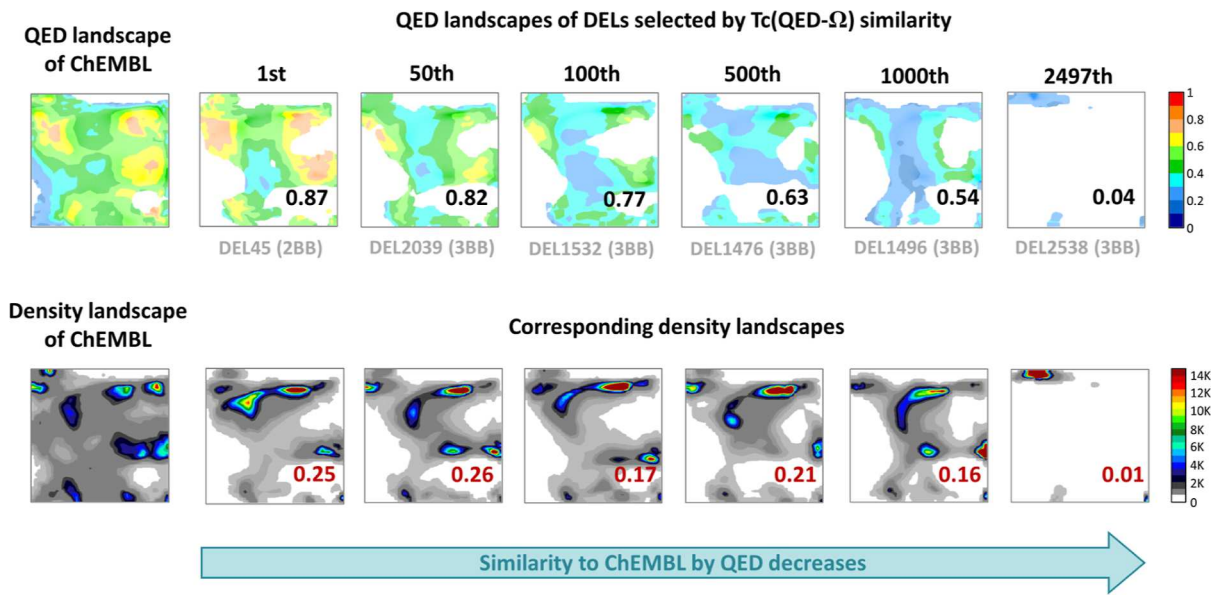


Figure 10. First row: On the left: QED landscape of filtered ChEMBL28. On the right: QED landscapes of DELs ranging from the most similar to the least similar to ChEMBL sorted by their Tanimoto coefficients calculated based on their QED-Ω with respect to ChEMBL (in black). Second row: On the left: density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

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

and chemically meaningful and allow adequate sorting of DELs 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

Data Availability Statement

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.

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.

■ 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/, accessed 2020.
(32) Enamine Ltd. https://enamine.net/, accessed 2020.

https://doi.org/10.1021/acs.jcim.3c00520
J. Chem. Inf. Model. 2023, 63, 4042−4055

Summary

The performance of the proposed in this work library vectorial representations was verified using several case studies. First, Φ, Λ, Γ, and Γw were calculated for the fully enumerated 88M DEL and its 1M-sized representative subset. Essentially, a library vector that can precisely encode the structural composition of the library should show high similarity between these two. Either a Tanimoto coefficient or coverage score was used to measure the similarity between the full library and its subset. In the case of Φ and Λ, the Tc were 0.99 and 0.98 confirming their ability to encode structural composition of the DEL in question. Coverage score based on Γ and Γw was 0.09 and 0.87, respectively. It means that the subset covers only 9% of chemotypes from the full library but they correspond to 87% of compounds present in the full collection. Hence, the subset lacks very rare chemotypes while covering popular ones from the full library. Overall, the present case study allowed to estimate the reliability of the proposed vectors and fingerprints in meaningfully encoding structural and property information of an entire compound collection.

Once the performance of the library representations was confirmed, 2.5K DELs were ranked by their similarity to ChEMBL28. The analysis of density landscapes revealed that Φ logically classified the most structurally diverse DELs (according to their chemical space span on the landscape) as most similar to ChEMBL. Whereas DELs concentrated in a single density peak, covering very focused regions of the chemical space, were classified as least similar. An in-depth analysis of the structural overlap between the most similar DEL and ChEMBL expectedly showed that the overlapping density peaks from both libraries contain similar compounds.

The Ω library vector enabled the ranking of DELs according to their similarity to ChEMBL based on both structural and property similarities. The properties used for the analysis included molecular weight (MW), logP, H-bond acceptor and donor counts, number of rotatable bonds, and the quantitative estimate of drug-likeness score (QED). The comparison by Ω enabled the immediate identification of the most similar and dissimilar DELs to ChEMBL based on various property distributions and chemical space coverage at once, thereby accelerating the multi-parameter library selection process. The proposed library vectors being GTM-based, offered a quick method for DEL comparison

based on structure and properties while simultaneously enabling the interpetability through detailed chemical space inspection using 2D maps.

6. Meta-GTM: a tool for Chemical Library Space visualization

Introduction

The space of DELs generated in our previous study contains 2.5K libraries. To be able to select a specific DEL from such a large number of libraries possessing the required structural coverage and property profile for a drug discovery project in hand, methods bypassing simple pairwise library comparison can be useful. For example, the full CLS depiction in one plot allowing to see the relationships between thousands of libraries can significantly speed up the preliminary library analysis and selection of candidate collections.

In our previous work, we introduced library descriptor vectors that define the position of compound collections within the multidimensional Chemical Library Space (CLS). In this way, the CLS of DELs was represented from different perspectives using various library encodings. The latter were then used to calculate the similarity between each DEL and ChEMBL, a database of biologically tested molecules. While pairwise comparison of libraries using these encodings can be effective, it does not provide information about the relationships between more than two libraries. However, understanding the relationships between all DELs in the CLS, as well as their relation to

Glossary

CLS – Chemical Library Space, which is a conceptual space defined by library descriptors. The latter define the position of the compound library in the CLS where it is considered an individual object.

GA – In the context of ML model optimization, a Genetic Algorithm is an evolutionary optimization technique that iteratively evolves a population of parameter sets (in the case of GTM - map parameters) to find the most optimal combination according to the selected scoring function through processes such as selection, crossover, and mutation.

Meta-GTM landscapes – These are meta-GTMs where each node on the map contains one or more compound libraries and where nodes can be colored based on various library characteristics.

ROC AUC – In virtual screening, ROC AUC is a model performance metric measuring its ability to correctly rank active compounds higher than inactive ones. The Area Under the ROC curve (AUC) plots the true positive rate against the false positive rate. A ROC AUC value of 1 indicates perfect discrimination between active and inactive compounds, while a value of 0.5 indicates no discrimination (random guessing).

ChEMBL, can be useful for reducing the number of libraries at the early stages of chemoinformatic analysis. This can save time for more detailed comparisons of only potentially useful collections for the task in question.

Just as GTM can visualize high-dimensional compound space, it can also provide a 2D map of the CLS. Hence, in this work, building on the success of meta-GTM in representing the compound vendor CLS as introduced by Gaspar et al.[50], we applied this approach to visualize the CLS formed by 2.5K DELs and ChEMBL. Meta-GTM is a GTM applied to reduce the dimensionality of the CLS defined by library descriptors. The prefix "meta" signifies that GTM is applied for a second time to visualize the CLS, but the principle remains the same. To identify the optimal meta-map parameters for visualization of each CLS defined by either Φ, Λ, or Ω a Genetic Algorithm (GA) optimization was used. The goal was to create a map that accurately visualizes the space of DELs, positioning them according to their similarity to ChEMBL and each other within the CLS defined by library vectors. To achieve this, the top 100 DELs most similar to ChEMBL in the CLS were selected using the Tanimoto coefficient based on their library vectors. The performance of the meta-GTM was evaluated by determining if these top 100 DELs remained in the 100 closest library list to ChEMBL on the meta-map. This was assessed by calculating the Euclidean distance between DEL-ChEMBL pairs on the map. Subsequently, the ROC AUC was computed, with the top 100 DELs in the initial CLS serving as positives and the remaining DELs as negatives. The best-performing meta-maps visualizing the CLS of DELs and ChEMBL (defined by either of Φ, Λ, or Ω) were associated with ROC AUC values ≥ 0.89. This means that all meta-maps created in this study allowed to consistently find the top 100 most similar DELs to ChEMBL from the initial CLS with the lowest Euclidean distances with respect to ChEMBL on the meta-maps. Hence, these meta-maps were declared to preserve the interlibrary neighborhood relationships observed in CLS.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

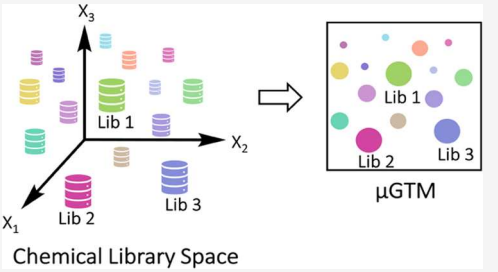
Meta-GTM: Visualization and Analysis of the Chemical Library Space

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

Cite This: J. Chem. Inf. Model. 2023, 63, 5571−5582 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In chemical library analysis, it may be useful to describe libraries as individual items rather than collections of compounds. This is particularly true for ultra-large noncherry-pickable compound mixtures, such as DNA-encoded libraries (DELs). In this sense, the chemical library space (CLS) is useful for the management of a portfolio of libraries, just like chemical space (CS) helps manage a portfolio of molecules. Several possible CLSs were previously defined using vectorial library representations obtained from generative topographic mapping (GTM). Given the steadily growing number of DEL designs, the CLS becomes "crowded" and requires analysis tools beyond pairwise library comparison. Therefore, herein, we investigate the cartography of CLS on meta-(μ)GTMs—"meta" to remind that these are maps of the CLS, itself based on responsibility vectors issued by regular CS GTMs. 2,5 K DELs and ChEMBL (reference) were projected on the μGTM, producing landscapes of library-specific properties. These describe both interlibrary similarity and intrinsic library characteristics in the same view, herewith facilitating the selection of the best project-specific libraries.



1. INTRODUCTION

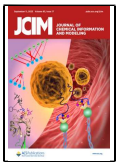
Historically, pharmaceutical companies have relied on in-house HTS libraries for hit identification. However, due to advances in disease characterization and the emergence of new challenging targets, it is becoming increasingly difficult to find suitable drug discovery starting points using this screening paradigm.[1] Hence, HTS is now complemented by other approaches such as DNA-encoded library (DEL) technology.[2] Since the latter allows for simple and fast combinatorial synthesis and screening of an entire library of DNA-encoded compounds (one DEL can contain up to trillions of compounds[3]), it enables the exploration of new areas of the chemical space in just a single screen, allowing for the identification of ligands for novel target classes or intractable target families.[4] However, it raises the problem of the appropriate library selection from a possible enormous chemical library space (CLS)[5] of thousands of DELs.

In our recent study,[5] we introduced CLS vectors encoding its structural or/and property information, derived from a generative topographic map of the corresponding compound collections. Herewith, the standard chemoinformatics of molecules can be seamlessly generalized to libraries. Interlibrary similarity/overlap scoring is the key to selecting analog or complementary libraries to a given compound collection. The next step of this generalization, covered in this work, is the cartography of CLS, allowing visualization of the similarity relationships of large numbers of considered chemical libraries, beyond simple pairwise estimation of their degree of overlap.

The visualization of the high-dimensional CLS obviously requires dimensionality reduction techniques. The most popular methods used for the chemical space visualization are principal component analysis[6] (PCA), self-organizing maps[7] (SOM), stochastic proximity embedding[8] (SPE), and t-distributed stochastic neighbor embedding[9] (t-SNE). However, GTM is advantageous due to its fuzzy nature, limiting the information loss upon dimensionality reduction, big data compatibility, ability to analyze new data without retraining, and ability to support (predictive) property landscapes.[10−12] Detailed comparison of GTM with some other dimensionality reduction techniques is reported in our review paper.[12]

A method that was recently used to specifically visualize the CLS and analyze the relationships between libraries is the chemical library network[13] (CLN), which is a generalization of chemical space networks[14] (CSNs). In this work, the authors represented a compound library as a collection of fingerprints and used an extended similarity index[15] to quantify the similarity between a pair of data sets. The libraries were represented as nodes of the CLN, whereas the connections between them were defined by the extended similarity index value. However, this approach has some limitations: the data set similarity represented by the extended similarity index cannot be explicitly chemically interpreted and, as the authors noted, the information about properties associated with the library is not implemented in CLN. In addition, the CLN was created for the CLS composed only of 19 libraries, and

Received: May 11, 2023
Published: August 21, 2023



https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

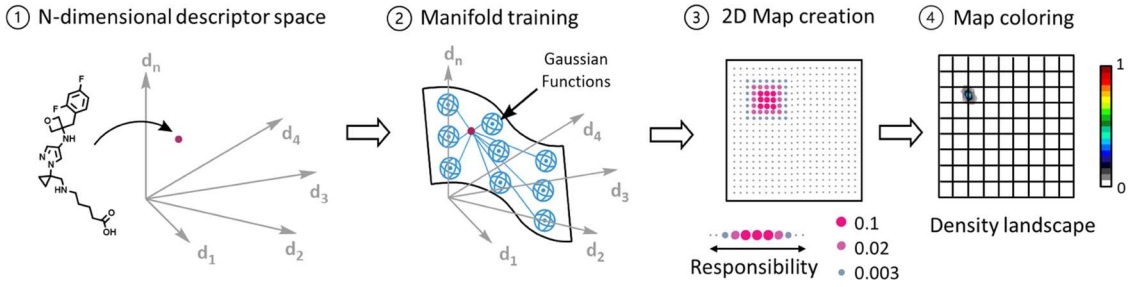


Figure 1. Generative topographic map (GTM) creation with the subsequent coloring of the map by the quantitative distribution of compounds in the chemical space, giving rise to the density GTM landscape.

network-style visualization techniques become notoriously unreadable when the number of mapped items exceeds 10^2.

In our previous work,[5] CLS defined as a high-dimensional vectorial space was based on the three types of below-mentioned GTM-based library descriptor vectors:

(1) Normalized (library size-independent) cumulated responsibility vector (Φ)—implicitly representing the "chemotype" distribution in the library (molecules of similar responsibility patterns tend to share common "chemotypes").

(2) Library-modulated cumulated responsibility vector (Λ)—weighing the library responsibility vector, by giving more importance to nodes known to harbor an external reference library.

(3) Property-modulated cumulated responsibility vector (Ω)—weighing the library responsibility vector by the (physicochemical, target binding, etc.) property values mapped on every node.

CLS is a regular vector space that can be visualized by any dimensionality reduction method. Here, it was done using GTM due to consistency reasons of using the same method that was employed for the library vector creation. Since the CLS descriptors are themselves based on the GTM projections of compound chemical space, the CLS maps will be further referred to as meta-GTMs (μGTMs).[16] In this work, we aimed to analyze the CLS populated by 2497 virtual DELs and ChEMBL using μGTM. ChEMBL was selected as a reference database due to its richness in molecules with high functional diversity (displaying a broad range of biological activities), thus being a good reference for the selection of diverse DEL that may be further used for primary screening. The main goal was to develop a μ-map of the CLS that will conveniently position DELs with respect to ChEMBL and each other according to their similarity observed in an N-dimensional space defined by library descriptor vectors. Overall, seven μ-maps were built on Φ, Λ, and Ω modulated by different properties (MW, log *P*, number of H-bond acceptors, number of H-bond donors, and quantitative estimate of drug-likeness, QED).

The power of this visualization technique was evidenced by generating various μGTM landscapes providing a clear overview of the ensemble of considered libraries. To create such landscapes, μ-maps were colored based on the intrinsic library features such as the number of chemistry cycles and reaction types (either coupling or heterocyclization transformation). This allowed us to instantaneously identify and highlight the nature of the closest DEL neighbors of ChEMBL: 2-building block (BB) or 3BB library, employing a particular reaction type. Maps can also be colored by the estimated cost of each DEL synthesis, substantially simplifying the process of selection of the library for screening for medicinal chemists considering multiple factors at the same time. This functionality of μGTM—the ability to color the map by any library feature—is useful to consider extra information when analyzing the CLS in addition to the already existing structural and/or property similarity information of the libraries.

Our results indicate that μGTM is a valuable tool for visualizing the CLS from different perspectives, being the first tool to date that supports property-sensitive visualization and allows us to include complementary library-related information in the analysis. Practically, this tool can facilitate decision-making for medicinal chemists when there is a problem with the selection of a single compound collection from a pool of thousands of libraries by considering not only chemical space coverage but also extra library parameters such as physicochemical profile, drug-likeness conformity, availability, facility of synthesis, cost, etc., all by providing comprehensible visualization of the CLS.

2. DATA

2.1. ChEMBL. The ChEMBL28 database containing molecules biologically tested against more than 15,000 targets was selected as a reference library.[17] Its standardization was performed in one of our previous works[18] in conformity with the protocol implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics at the University of Strasbourg, using the ChemAxon Standardizer.[19] It was then filtered according to the rules of DEL-likeness derived in our previous work.[5] Hence, until otherwise stated, in the text, we always refer to the filtered ChEMBL28 containing 1,605,370 molecules. After standardization and filtering, ISIDA fragment descriptors of type *IA-FF-FC-AP-2-3* (ISIDA fragment sequences of 2 and 3 atoms labeled by their CVFF force field types and formal charge using all paths) were calculated.[20] These descriptors were used since they were earlier employed for the universal map #1 creation that we use for each library chemical space visualization in this study.[21]

2.2. DNA-Encoded Libraries. 2497 virtual DELs used in this work were generated in our previous study[18] using eDesigner[22] for DEL design and enumeration. They were designed using 79 K building blocks (BBs) from eMolecules and Enamine suppliers and DEL-compatible reactions encoded in eDesigner. Only 1M representative subset of compounds per DEL was enumerated, although the designed DEL sizes varied from 1M to 6.7B. All 2497 DELs were standardized according to the previously described procedure that was used for ChEMBL, and the same descriptors were calculated.

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

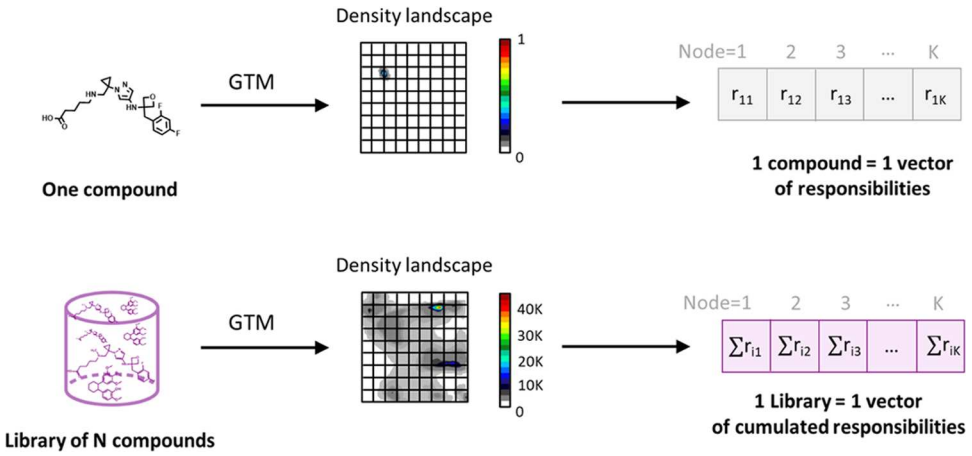


Figure 2. Scheme describing how a responsibility vector is obtained for a single molecule using GTM (top) and how a **CRV** is obtained for the whole library of compounds (bottom).

3. METHODS

3.1. Visualization of the CLS on a μGTM. *3.1.1. Generative Topographic Mapping.* GTM is an unsupervised dimensionality reduction method that consists of transforming the initial N-dimensional space into a two-dimensional (2D) latent space by fitting a 2D flexible manifold (rectangular hypersurface) to the data space (Figure 1, steps 1 and 2). The GTM algorithm adjusts the degrees of freedom by controlling the manifold shape in order to maximize its proximity to all "frame items" used to outline the relevant space zone. The molecules (data points) can be projected to the optimized manifold with node-specific probabilities termed responsibilities. Then, the manifold is relaxed back to its planar form to give a 2D map where the position of each molecule is defined by a responsibility vector (Figure 1, step 3). The map can be colored based on the cumulated responsibility to give a density landscape (Figure 1, step 4) or by some property, giving rise to property landscapes. The relevance of the latter, as predictors of properties of newly projected items, may serve as quantitative estimators of the quality of the built map.

The flexibility of the manifold is tunable and is controlled by the number of radial basis functions (RBFs); the higher it is, the more flexible the manifold and thus better its ability to adapt irregular, nonlinear shapes to describe the data.[11] In addition, the map size, the RBF width factor, and regularization coefficient complete the set of map-defining hyperparameters. These must be tuned in order to maximize map quality or fitness[23]—here, by a genetic algorithm (GA).

The GTM that was used in our previous study[5] for visualization of each library space and further for the generation of library-representative vectors is the "universal" GTM #1, which was developed by Casciuc et al.[21] This map was trained to be able to robustly separate active from inactive compounds over a vast set of diverse biological targets. μGTM tuning followed a similar procedure, except that the specific fitness criteria (measuring the preservation on the μGTM of interlibrary neighborhood relationships in observed CLS) needed to be introduced. The detailed optimization procedure is described in Section 3.1.3.

3.1.2. Library Descriptor Vectors. A single compound on the GTM is described by a responsibility vector, outlining the fuzzy levels of association of the compound to each map node. To encode a whole library by a single vector, the sum of responsibility values over all molecules of the library in each node of the GTM can be calculated to give a cumulated responsibility vector (**CRV**) (see Figure 2), described in more detail in our previous work.[5] Since GTM nodes are predominantly associated with specific "chemotypes" (common structural traits—scaffold, pharmacophore, size, etc.), the **CRV** allows us to quantify the occurrences of such chemotypes in a library.[5]

To obtain a library size-independent library descriptor vector Φ, **CRV** values can be normalized[5] by the library size *N*. To stress the importance of some nodes of the map, such as those densely populated in the reference compound collection, **CRV** can be modulated with respect to the compound distribution of another reference collection to give library-modulated **CRV** (Λ). For a viewpoint integrating compound property information, the responsibility values may be weighted by the property values of those molecules.[5] This defines property-modulated **CRV** (Ω). Formulas used to calculate Φ, Λ, and Ω values are given in the equations section of the Supporting Information (SI).

3.1.3. μGTM Creation. The terminology μGTM[16] is used to highlight that the input CLS vectors are themselves originating from standard chemical space GTM. As already mentioned, a GA was used to search an optimal parameter set producing a μ-map that will optimally visualize the CLS. This, of course, implies a quantitative definition of "optimality"—a CLS map fitness score. Typically, the fitness of regular CS maps, hosting individual compounds, is related to the propensity of the map to regroup molecules of similar properties and avoid colocalization of compounds with different properties: neighborhood behavior (NB) compliance. This is possible only based on "training" sets of compounds of experimentally known properties. In the CLS context, μ-mapped items are libraries and there are no properties measured to globally characterize each library as an object. Therefore, μ-map fitness was redefined in terms of the quality of preservation of interlibrary distances as calculated in the initial CLS descriptor space. We aimed to select a μ-map where the top 100 closest DELs to ChEMBL in the CLS vector space will be preferentially rendered as the top 100 closest neighbors of ChEMBL on the map. First, the list of the top 100 nearest DEL neighbors of ChEMBL is established by computing all ChEMBL-DEL Tanimoto similarity scores in the corresponding descriptor space (see the equations section in the SI) and sorting them accordingly (note—any DELs immediately

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

beyond the 100th position in the list but having an overlap score equal to the 100th ranked DEL are also included within this "top 100" list, for consistency). After the manifold is fitted according to the current μ-map parameter set (GA chromosome), ChEMBL and DELs are projected on the manifold, and latent space coordinates (x, y) are calculated for each library (as geometric centers of their responsibility clouds). All DELs are then ranked with respect to their increasing latent space Euclidean distances $\sqrt{(x_{\mathrm{ChEMBL}} - x_{\mathrm{DEL}})^2 + (y_{\mathrm{ChEMBL}} - y_{\mathrm{DEL}})^2}$ with respect to ChEMBL. Based on this ranking, a receiver operating characteristic (ROC) can be plotted considering DELs of the top 100 CLS neighbor set as the "positives" and all others as "negatives". If the top 100 neighbors are consistently found among those with the lowest Euclidean distances, this translates into a high ROC area under curve (AUC) value. Thus, this ROC AUC was used as the μGTM fitness score. The best μ-maps here were selected from the list of five top-scoring GA chromosomes all with ROC AUC $\geq$ 0.89.

The full list of μGTM parameters that were optimized by GA is given in Table S3 in the SI. The maximal number of parameter configurations to be explored was set to 5000, and the maximal number of generations indicating when to stop if no improvement of so far fittest solution is observed was set to 1000. The GA optimization duration typically lasted 1−5 days (in all cases, an optimal solution was already found after 1 day; additional calculations were increasing the scoring function value only by 0.01 for new chromosomes). Separate GAs were launched on 16 and 48 CPU machines (Intel Xeon W-2145 and Intel Xeon Silver 4214(R)) since the creation of seven meta-maps necessitated the launch of seven optimizations.

Since we used different types of library descriptor vectors, several μGTMs were created. The aim was not to select the best descriptor type; thus, CLS vectors did not compete against each other. Instead, one map per vector type was created. Φ- and Λ-based μ-maps were used to analyze the CLS from the library structural similarity point of view. On Ω-based μ-maps, the CLS was analyzed by considering both the property and structural similarity of libraries. Five properties were selected for analysis: MW, log P, number of H-bond donors, number of H-bond acceptors, and quantitative estimate of drug-likeness (QED).

Like with any GTM, a μGTM landscape can be colored based on projected properties, starting with basic cumulated responsibility (some CLS zones being more "crowded" in terms of resident libraries than others), library "status" (being a member of top 100 similar libraries relative to some reference database, for example), or other characteristics (underlying reaction type, number of building blocks engaged in the library compounds, etc.). Two alternative ways to display landscapes—"pixelized" and "continuous"—were both employed in this work. In "pixelized" landscapes, only the state of individual nodes is shown (the relative populations of classes, or the average property of the node, rendered as a square, encodes its color, whereas color intensity/transparency reflects the overall node population size, i.e., cumulated responsibility). In "continuous" landscapes, the nodes are not emphasized, and the landscape is obtained by the interpolation of the properties of neighboring nodes.

4. RESULTS AND DISCUSSION

4.1. Neighborhood Preservation on the μGTM.

Figure 3 shows the μGTM fuzzy class landscapes visualizing the CLS formed by the pool of 2497 DELs and ChEMBL built on Φ (left) and Λ (right) library descriptor vectors. Landscapes are colored by the relative occurrence of the top 100 DELs (class "red") among all other DELs (class "black") residing in each zone. Red nodes contain DELs inside the "top 100" library list, black ones contain DELs outside this list, and other nodes contain both library classes. For clarity, ChEMBL is represented as a cross on the node closest to the center of mass of its responsibility distribution.

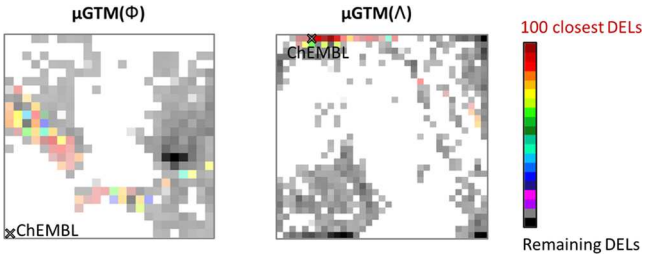


Figure 3. μGTM class landscapes built on Φ (normalized cumulated responsibility vector) (left) and ChEMBL-modulated Λ (ChEMBL-modulated cumulated responsibility vector) (right) of the pool of 2497 DELs and ChEMBL. The latter is the reference item (not included in the pool of mapped libraries) but positioned in terms of its (x, y) coordinates, i.e., center of its responsibility distribution. Landscapes are colored by the relative occurrence (see the color scale) of top 100 DELs (class "red") among all other DELs (class "black") residing in each zone. Map parameters are given in Table S2 of the SI.

On the μGTM(Φ), ChEMBL is located quite far from any DELs, with the top 100 closest libraries (T_c = 0.27−0.38 in Φ-based CLS) situated on the "shores" of the DEL islands nearest to ChEMBL. Hence, the top 100 DELs indeed stay closest to ChEMBL on μGTM(Φ). This is in agreement with the high fitness score (ROC AUC = 0.93) of this μ-map. On the μGTM built on Λ, the majority of the closest 100 DELs (T_c = 0.9−0.91) are located quite near to ChEMBL and some of them even overlap—consistently with the high ROC AUC = 0.94 of this μ-map. NB compliance of μ-maps is, in this sense, clearly achieved.

For a more detailed NB analysis on the μGTM(Φ), the DELs residing (with responsibility >0.8) in two nodes—one of the closest and one of the farthest to ChEMBL, respectively—were selected. Their chemical spaces were visualized by mapping on the "universal map" #1 of biorelevant chemical space[21] (see Figure 4). The closest node contains DEL3667, which is based on widespread coupling reactions: carboxylic acid/amine condensation, aldehyde reductive amination, and guanidinylation (see Figure 5). Its full size is 95.6M compounds, which are based on some highly popular building block classes (13 formyl carboxylic acids, 9 611 and 765 aliphatic amines). The farthest node comprises 104 DELs, most of them employing exclusively heterocyclization reactions, among which DEL2266. This DEL is based on 278 haloaryl carboxylic acids, 23 azides with an additional amine group, and 213 α-bromo ketones, giving rise to 1.4M compounds (Figure 5).

From their density landscapes in Figure 4, it appears that DEL3667 and ChEMBL are similar—with most of their

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

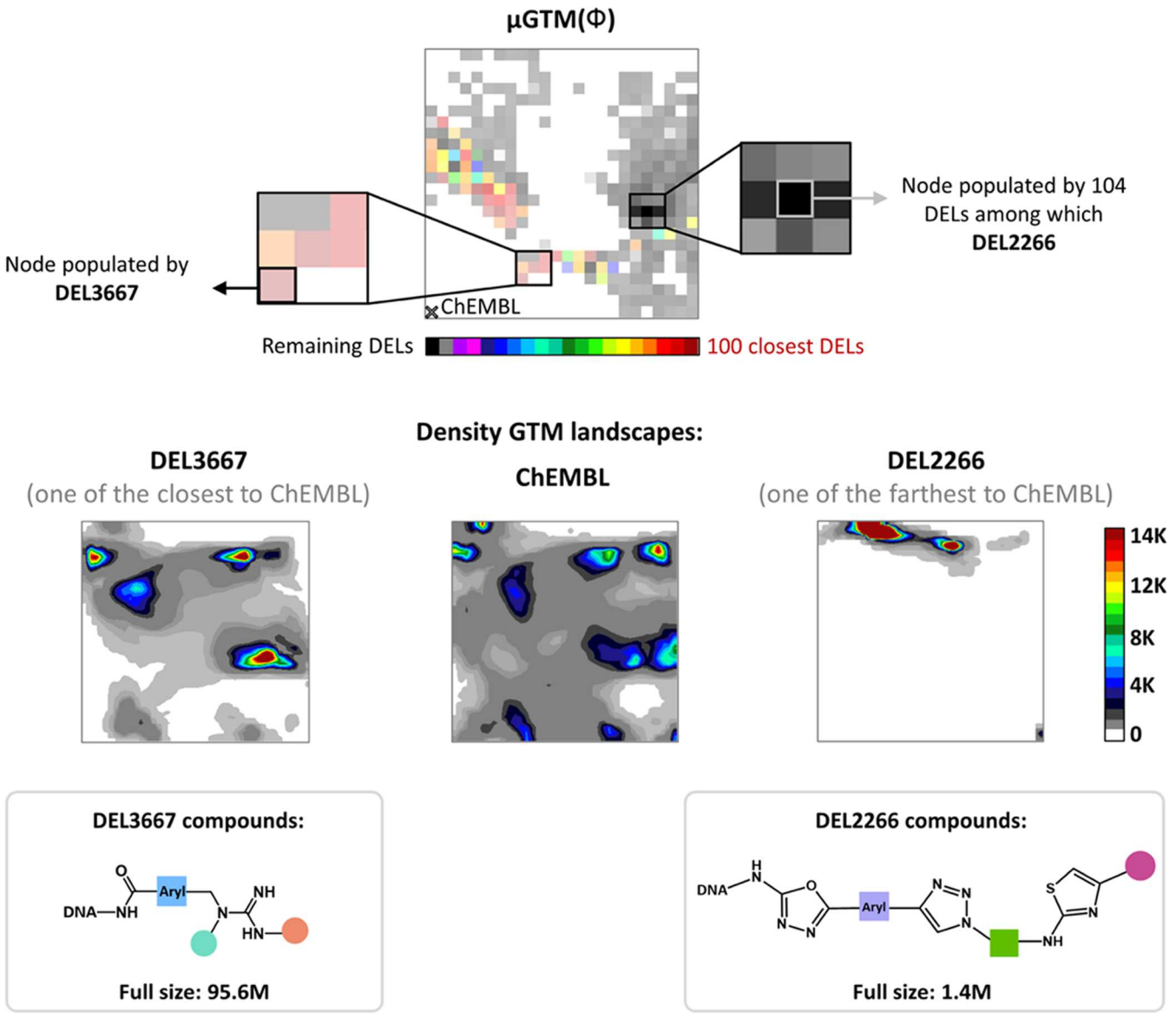


Figure 4. Top: class μGTM(Φ) landscape colored by the relative occurrence (see the color scale) of top 100 DELs (class "red") among all other DELs (class "black") residing in each zone. The brightness is used to differentiate highly populated nodes from scarcely populated ones. Bottom: density landscapes of DEL3667, ChEMBL, and DEL2266. The schemes depicting the skeleton of compounds present in each DEL and their full sizes are given below the corresponding density landscapes. Each map's parameters are given in Table S2 of the SI.

density peaks, representing the most populated zones of the chemical space, coinciding. This DEL is, expectedly, closer to ChEMBL in CLS than DEL2266 (see Tanimoto scores in Table 1), as the μ-maps were specifically selected to respect the initial CLS neighborhood. This example should not be read as a "validation" but as an illustration of how this framework of maps may intuitively convey an understanding of the relationships between libraries, beyond abstract Tanimoto scores that cannot be interpreted out of context. The density landscape of DEL2266 in turn is distinct from ChEMBL, with their density peaks not overlapping, which indicates their dissimilarity in terms of the covered chemical space. This is because of the lack of diverse BBs accessible for only heterocyclization-based DEL synthesis—the case of DEL2266. A more detailed look at these DELs is given in Figure 5 where each chemistry cycle along with the number of BBs that were used for DEL enumeration is shown.

To verify whether the neighborhood relationships between DELs are preserved on the μGTM, as an example, DEL1847 (the library with the highest similarity to ChEMBL in Φ-based CLS) was taken and its T_c similarity relative to all other DELs was calculated. Nine most similar DELs with $T_c \geq 0.7$ were selected, and they were spotted on the map. As seen from the μGTM in Figure 6, all of them are located either in the same node as DEL1847 or in neighboring nodes (colored in green), as expected. Unsurprisingly, DELs of a similar chemical space coverage project close to each other on the μ-map. Likewise, nine least similar DELs to DEL1847 were selected out of the 22 equally dissimilar DELs, with $T_c < 0.01$. As seen in Figure 6, these DELs are concentrated in the two nodes (colored in red) distant from the node with DEL1847.

The density landscapes in Figure 6 show that expectedly, the nine libraries closest to DEL1847 have highly similar density landscapes. While the density landscapes of the farthest DELs completely differ from the landscape of DEL1847, all of them

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

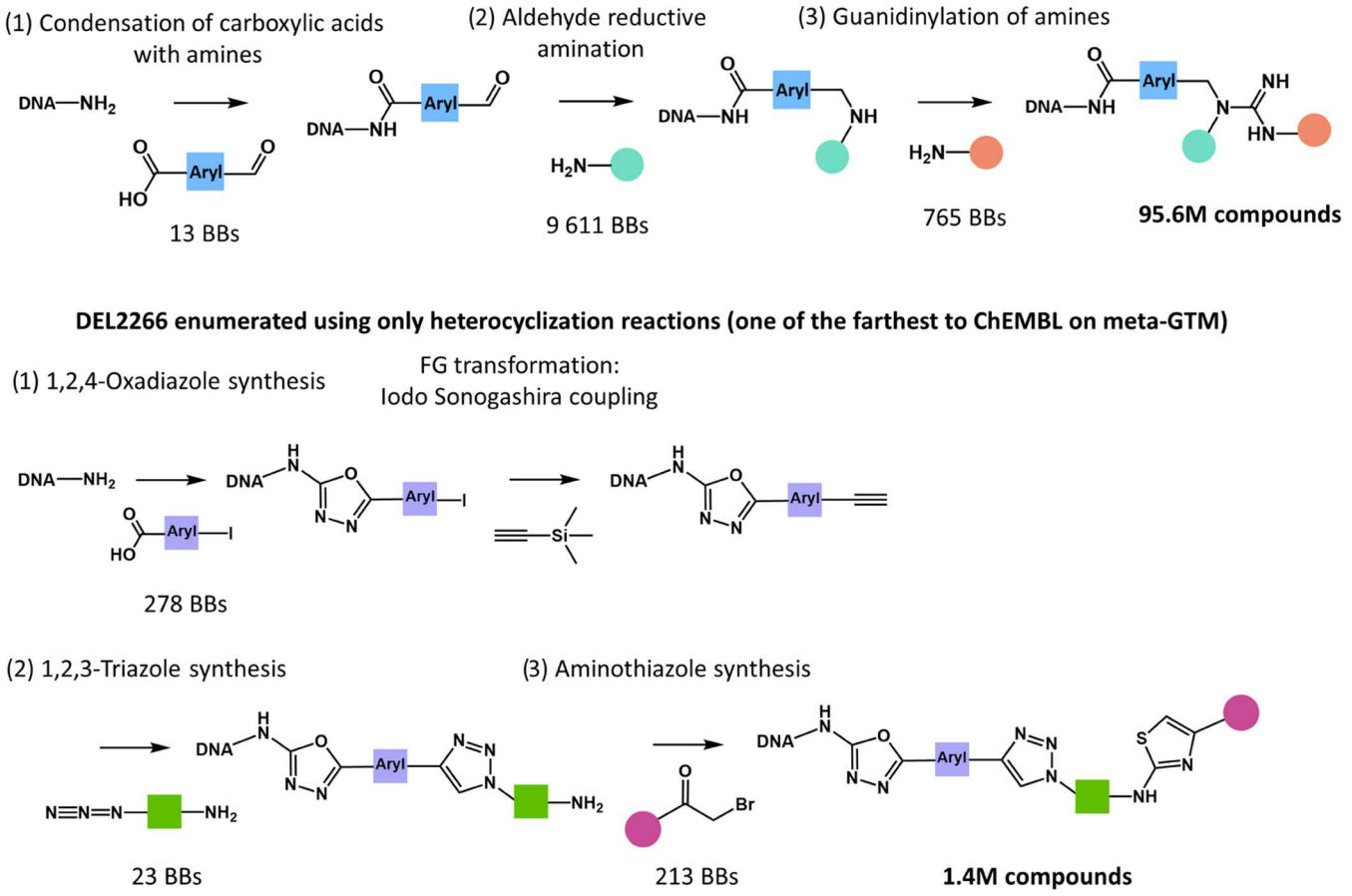


Figure 5. Examples of DELs enumerated by either only coupling reactions (DEL3667 located closest to ChEMBL on μGTM) or heterocyclizations (DEL2266 located farthest to ChEMBL on μGTM) using commercially available building blocks. For simplicity reasons, DNA-encoding steps are omitted. Additional reagents and reaction names are not given; only the building blocks and reaction types are shown according to the information present in the SI of eDesigner.[22]

Table 1. Tanimoto Values Calculated Based on Either Φ or Λ for DEL3667 and DEL2266 with Respect to ChEMBL

DEL id	T_c (Φ)	T_c (Λ)
DEL3667	0.33	0.91
DEL2266	0.003	0.81

cover the same, focused area of the chemical space. However, minimal T_c (maximal distance in the CLS) does not need to translate to the maximal Euclidean separation on the map because the CLS versus latent space distance relationship need not be linear and not even monotonic. NB compliance simply implies that no "remote" library in CLS is projected among the nearest neighbors of a reference library.

4.2. Analysis of μGTM Landscapes. *4.2.1. μGTM Landscapes Colored by the Number of Building Blocks (Chemistry Cycles) Used for DEL Synthesis.* DELs significantly vary with respect to the number of building blocks used for their synthesis, 2BB libraries being easier and cheaper to produce as well as being more Lipinski-rule compliant. In this regard, it is interesting to analyze DELs by the number of BBs incorporated in the final compounds on the μGTM landscape and spot, at a glance, how 2BB and 3BB libraries cohabitate (or not) in CLS, and how they are positioned with respect to a reference library. The μGTM landscapes are created by coloring the map by property value (here, #BB) of each item (here: the DEL, defined by its responsibility vector) on the map (see Figure 7). Since in the current work, #BB may be only 2 or 3, these landscapes are nothing but fuzzy two-class landscapes, where CLS zones occupied by only 2BB or, respectively, 3BB libraries adopt "extreme" colors and zones occupied by the two types of DELs translate the relative population onto the provided color spectrum. Overall, out of 2497 DELs, 97 are 2BB and 2400 are 3BB libraries.

In Figure 7, both μGTM class landscapes show that the closest libraries to ChEMBL are primarily 3BB DELs, while the majority of 2BB DELs are more widely distributed. This observation can be attributed to the incorporation of a third BB in DEL compounds, which inherently enables access to greater structural diversity. However, it should be noted that the number of 2BB libraries is almost 25 times lower than that of 3BB DELs. Therefore, it cannot be claimed that covering the same chemical space achievable by 3BB libraries using 2BB DELs is fundamentally impossible (after all, any BB may itself be envisaged as a coupling product of two precursor BBs—classifying DELs by the BB number only makes sense within the rather arbitrary context of a given BB pool). The class landscape also shows that there are very few 2BB-specific zones, whereas there are many zones that are 3BB-specific (the number of completely red zones is limited, whereas there is plenty of fully black ones). In other words, there is no specific diversity "niche" targetable only by generated here 2BB DELs, but there is plenty of chemical space that requires 3BB strategies—everywhere, not only around ChEMBL. It is

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

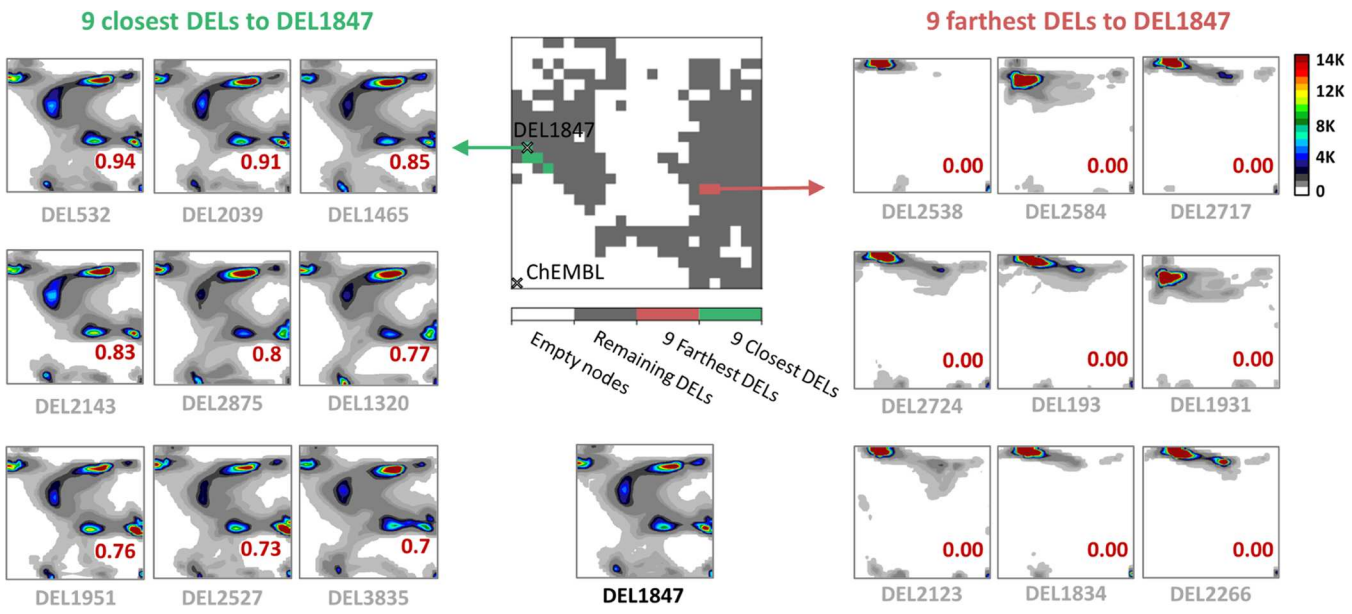


Figure 6. Middle top: μGTM(Φ) class landscape of 2497 DELs and ChEMBL. Zones of the map are colored based on the presence of DELs of a particular class in them: either in the nine closest library list (green) or the farthest library list (red) relative to DEL1847 or in the remaining library list (dark-gray). Middle bottom: density landscape of DEL1847. Left: density landscapes of the nine closest DELs to DEL1847 lying predominantly (with the probability to reside in the node >0.9) in the nodes of μGTM(Φ) colored in green. Right: density landscapes of nine farthest DELs to DEL1847 lying predominantly (with the probability to reside in the node >0.9) in the nodes μGTM(Φ) colored in red. For coherency reasons, all density landscapes are rendered using the ChEMBL density scale.

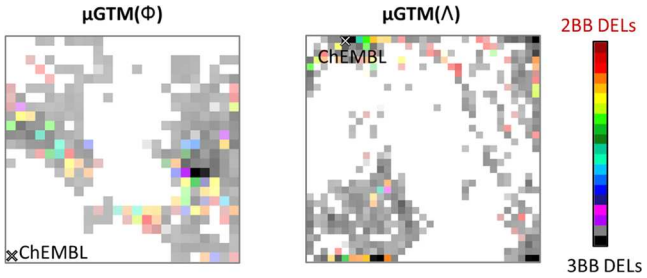


Figure 7. μGTM class landscape built either on Φ (normalized cumulated responsibility vector) or ChEMBL-modulated Λ (ChEMBL-modulated cumulated responsibility vector) of 2497 DELs, monitoring 2BB versus 3BB DELs. Map parameters are given in Table S2 of the **SI**. Out of 2497 DELs, 97 of them are 2BB and 2400 are 3BB libraries.

important to note that these observations are limited to the generated space of 2497 DELs in this work.

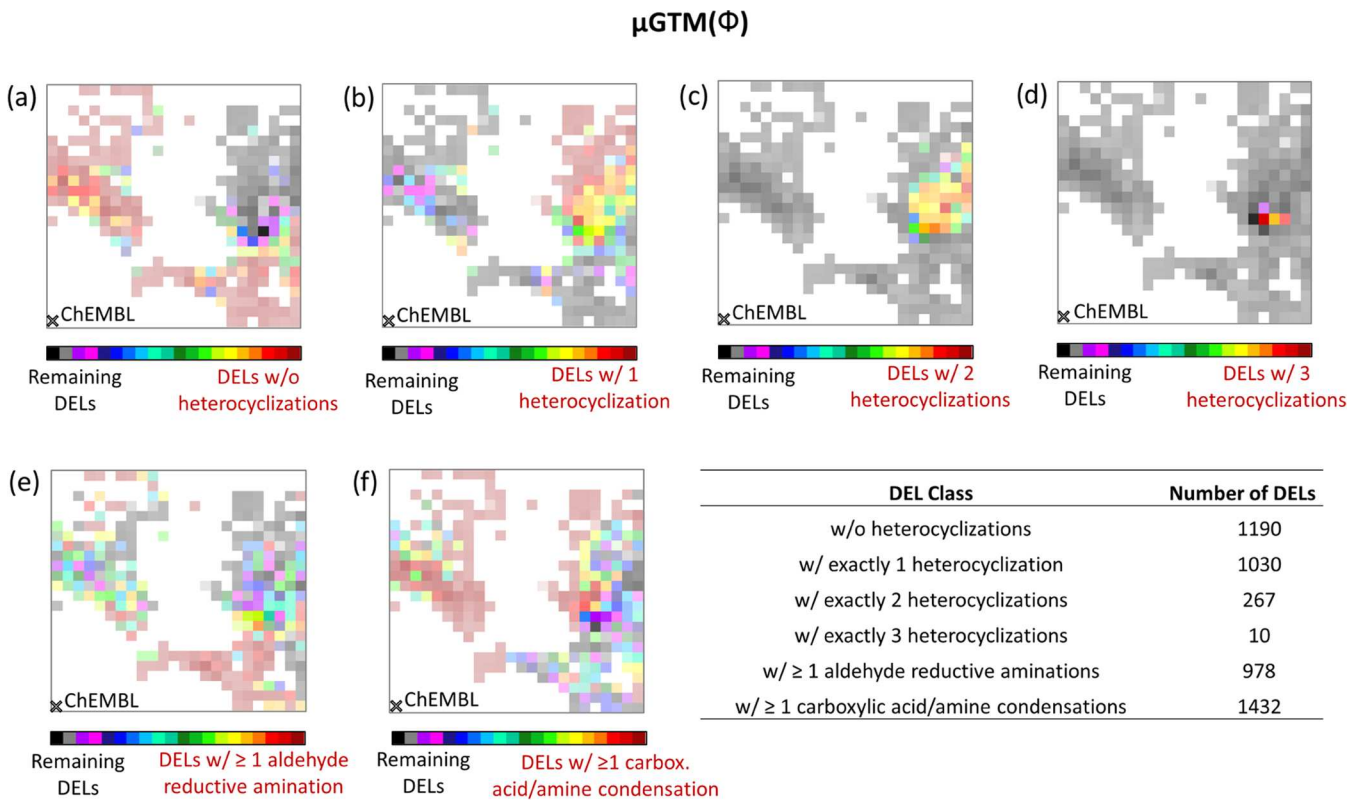
Altogether, based on the BB landscapes here, the number of BBs is not that distinctive for the chemical space covered by a DEL, unlike the reactions behind them. There are 2BB libraries having higher similarity to ChEMBL than 3BB DELs due to the use of BB-class abundant coupling reactions, e.g., DEL222 based on carboxylic acid/amine condensation and Ullmann-type N-aryl coupling with T_c = 0.25 in Φ-based CLS. There are also plenty of 3BB DELs displaying limited diversity and thus lower similarity to ChEMBL, for example, DEL3703 with $T_c(\Phi)$ = 0.01 based purely on heterocyclization reactions: oxadiazole, triazole, and aminothiazole syntheses. The difference in the ChEMBL coverage by these two DELs can be clearly seen by comparing their density landscapes to ChEMBL in Figure S1 of the SI.

4.2.2. μGTM Colored Based on the Reaction Types Used for DEL Synthesis. The μGTMs(Φ) (Figure 8) were colored according to the relative proportion of node-resident DELs based on a chosen reaction type[24] (coupling and heterocyclization, respectively) in their chemistry cycles. These are again fuzzy 2-class landscapes: DELs based on the given reaction type versus all others. The number of libraries corresponding to each class is given in the table in Figure 8. The complete list of reactions included in each type as encoded in the eDesigner tool used for DEL design and enumeration is given in Table S1 of the SI. For more information on the used reactions, see the SI of the corresponding paper.[22]

The landscape from Figure 8a is dominated by red since 1190 libraries out of 2497 DELs do not employ any heterocyclizations. This part of the generated DEL space can be attained using only coupling reactions, allowing for higher diversity due to the abundance of the available BBs.[25] In the case of heterocyclization-based DELs, the compounds are limited to a small number of scaffolds due to the introduction of the same heterocycle to final molecules and the lower number of BBs available for this type of transformation. Each heterocyclization in a way has an intrinsic scaffold bias that typically leads to a lower scaffold diversity in the resulting compounds than in the case of coupling reactions.

From the μGTMs in Figure 8 from (b) to (d), it appears that DELs forming one heterocycle upon synthesis are, however, exploring novel patterns of chemical space occupancy, represented by the red and intermediate-colored landscape areas, which are not reachable with simpler, pure coupling-based DELs. However, the closure of two or three heterocycles no longer pushes the respective DELs into the unchartered μ-map territory—they tend to focus toward the center of the mono-heterocyclization DEL space. Purely heterocyclization-based 3BB DELs on the landscape (d) in Figure 8 are all clustered in a small area of the CLS, showing their bias toward a particular chemical space region.

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582



DEL Class	Number of DELs
w/o heterocyclizations	1190
w/ exactly 1 heterocyclization	1030
w/ exactly 2 heterocyclizations	267
w/ exactly 3 heterocyclizations	10
w/ ≥ 1 aldehyde reductive aminations	978
w/ ≥ 1 carboxylic acid/amine condensations	1432

Figure 8. μGTM(Φ) class landscapes of 2497 DELs and ChEMBL and table with the number of libraries corresponding to each class. The maps are colored based on the ratio of DELs in the node having a particular reaction type in the chemistry cycles for their synthesis. Map parameters are given in Table S2 of the SI.

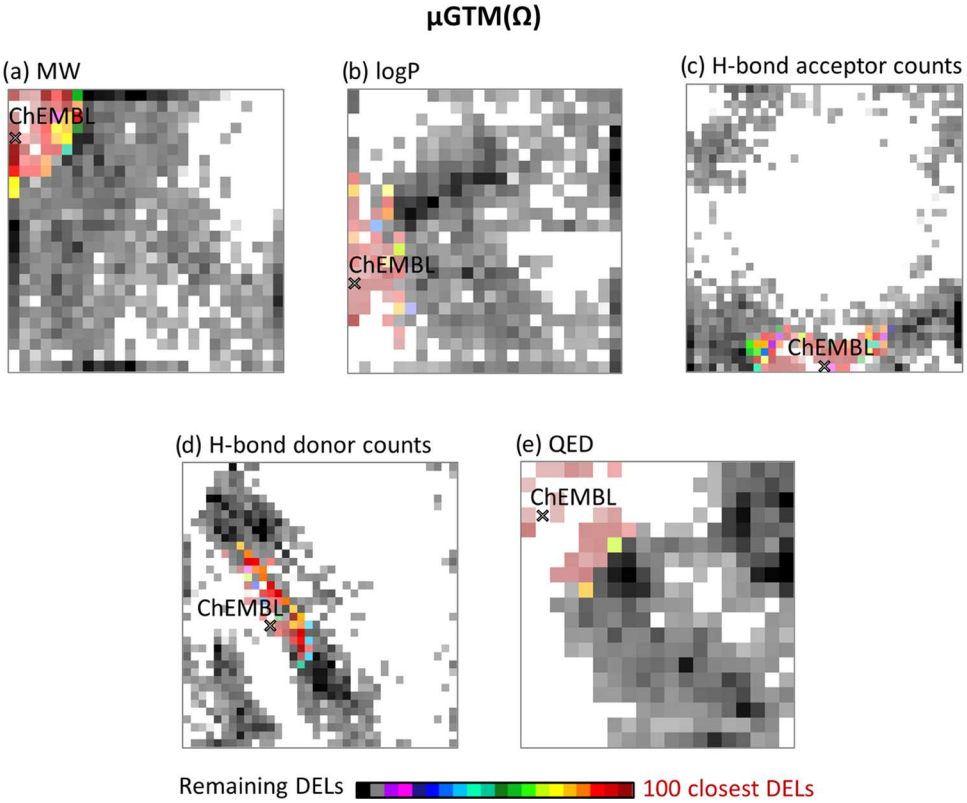


Figure 9. Class μGTM landscapes built on (a) MW-Ω, (b) log *P*-Ω, (c) H-bond donor counts-Ω, (d) H-bond acceptor counts-Ω, and (e) quantitative estimate of drug-likeness-Ω. Nodes are colored based on the relative fraction of the top 100 DELs among the residents. Map parameters are given in Table S2 of the SI.

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

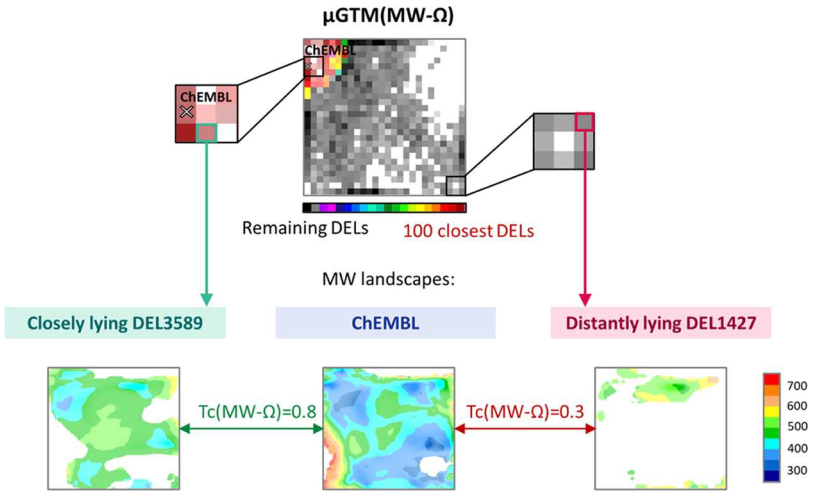
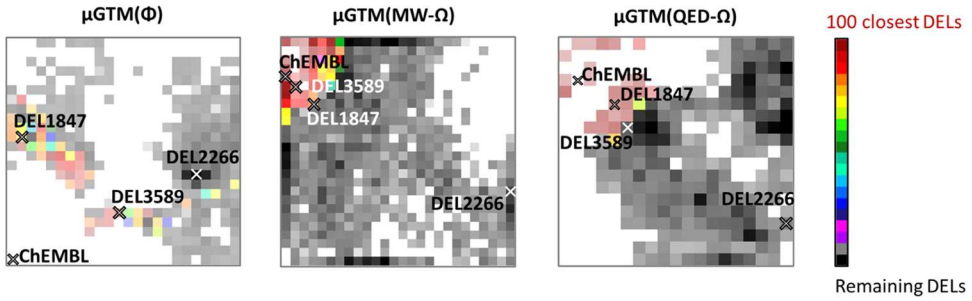


Figure 10. Top: μGTM(MW-Ω) class landscape where one of the closest and one of the farthest zones relative to ChEMBL is framed in black rectangles; ChEMBL is depicted as a cross. Bottom: MW landscapes of the DELs (DEL3589 and DEL1427) lying in the closest or farthest nodes, respectively. Each map's parameters are given in Table S2 of the SI.



	CLS(Φ)		CLS(MW-Ω)		CLS(QED-Ω)	
DEL id	Rank	Tc	Rank	Tc	Rank	Tc
DEL1847	1	0.38	222	0.75	32	0.83
DEL3589	34	0.3	6	0.8	130	0.75
DEL2266	2497	0.003	2493	0.17	2496	0.05

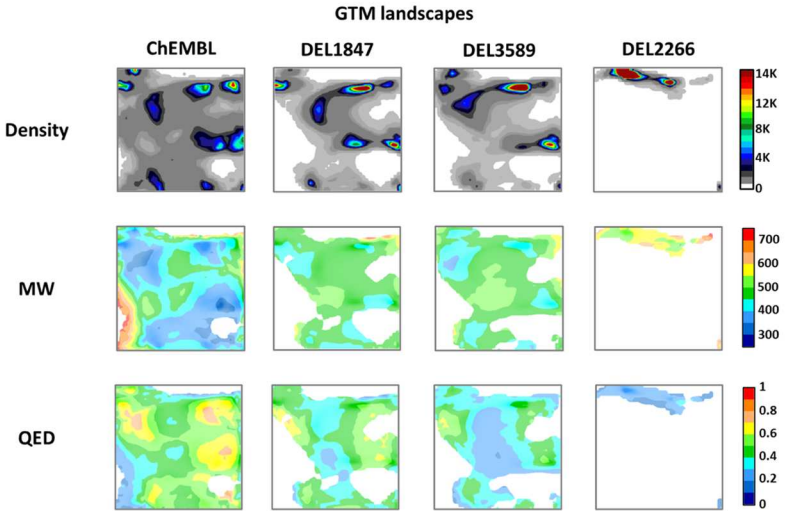


Figure 11. Class μGTM landscapes built on (1) Φ, (2) MW-Ω, and (3) QED-Ω. Nodes are colored based on the relative fraction of the top 100 DELs among the residents. The nodes where DEL1847, DEL3589, and DEL2266 reside (with the responsibility of ≥0.7) are marked by crosses. ChEMBL is positioned in terms of its (x, y) coordinates, i.e., center of its responsibility distribution, and shown as a cross as well. The table below contains the rank and T_c similarity values in the corresponding CLS, defined by either Φ, MW-Ω, or QED-Ω. The corresponding density, MW, and QED GTM landscapes are given below. For all landscapes, the same color scale corresponding to the density or property distribution of ChEMBL was used.

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

In Figure 8e,f, the two μGTMs are class-colored by the underlying reaction: aldehyde reductive amination (versus all other reactions) and condensation of carboxylic acids with amines (*idem*). On both of these maps, a quite broad area of the CLS is covered by DELs based on these reactions—the majority of the map nodes are colored in either red (resident DELs are all products of a given reaction) or intermediate colors (at least some resident DELs are using that reaction) but rarely in black. In other words, the state-of-art coverage of the CLS as achieved by the herein-studied 2497 DELs is impossible without involving reductive amination or amide formation as at least one synthetic step. DELs employing these reactions are also located close to ChEMBL in the CLS as can be seen from the maps—red zones are located near ChEMBL. This result is not surprising since for aldehyde reductive amination and condensation of carboxylic acids with amines, which are the most common coupling reactions employed in the DEL synthesis, a high number of structurally diverse building blocks exist.[18,24]

4.3. Property μGTMs. μGTMs were created for 2497 DELs and ChEMBL using MW, log *P*, number of H-bond acceptors, number of H-bond donors, and QED score-modulated Ω vectors as library descriptors (see Figure 9). As before, the produced μGTMs were represented as class landscapes where the first class (red) corresponds to the 100 DELs closest to ChEMBL in CLS (according to the Tanimoto similarity coefficient computed with corresponding Ω), whereas the second class (black) represents the remaining DELs. All maps in Figure 9 show that the ChEMBL is surrounded by or even overlaps with some of the 100 closest DELs, as expected.

To have a closer look at the property space spanned by the DELs positioned in the closest and farthest nodes of the property μGTM landscape relative to ChEMBL, their GTM property landscapes were analyzed. A μGTM built on MW-modulated Ω was taken as an example for such analysis (see Figure 10). One of the nearest libraries to ChEMBL on μGTM(MW-Ω) is DEL3589, also nearest in CLS, with T_c(MW-Ω) = 0.8. It is an 80M DEL enumerated using two aldehyde reductive aminations and Ullmann-type N-aryl coupling of arylhalides and amines. DEL1427, one of the remotest DELs relative to ChEMBL on μGTM(MW-Ω) and accordingly in CLS, with T_c(MW-Ω) = 0.3, is a 1.6M library based on guanidinylation of amines and two heterocyclizations—triazole and imidazole syntheses.

MW-colored GTM landscapes of these two DELs show the drastic difference between them in terms of property and chemical space coverage as well as in terms of similarity to ChEMBL. The landscape of DEL3589 covers similar chemical space as ChEMBL, and many zones on the two maps are colored likewise. The former means that DEL3589 reproduces well the structural distribution of ChEMBL and the latter—that the mean property values in the corresponding nodes match, confirming the similarity of this DEL to ChEMBL in terms of MW coverage. By contrast, the DEL1427 MW landscape does not correspond to the coloring of the ChEMBL landscape. Having DEL3589 and DEL1427 at the "antipodes" of the μGTM(MW-Ω) in Figure 10 makes perfect sense and can be easily grasped from GTM landscapes.

In order to analyze whether DEL positions relative to ChEMBL are different from one property CLS to another, the two most ChEMBL-similar DEL1847 and DEL3589 and the most dissimilar DEL2266 in Φ-based CLS were monitored on μGTM(MW-Ω) and μGTM(QED-Ω) (see Figure 11). Overall, it is visible that the neighborhood of ChEMBL and the DELs in question is affected by the property used for the μGTM creation. For example, the closest DEL1847 is not consistently in the top 100 neighbor list—on μGTM(Φ) and μGTM(QED-Ω), it is positioned in the top 100 red region of the μGTMs, but not on μGTM(MW-Ω)—it is ranked 222nd and thus positioned in the mixed zone of the corresponding μ-map. Coherently, on the density and property landscapes in Figure 11, DEL1847 has a QED score and quantitative compound distributions similar to ChEMBL. DEL3589 in its turn falls into the top 100 in the MW-Ω-defined CLS, which is visible from the μGTM(MW-Ω) as well as from its MW landscape—it is positioned in the red node on the μ-map and its MW distribution coincides better with ChEMBL than one of DEL1847, which is reflected by the higher coloration match in many zones of the MW landscapes. Naturally, both DEL1847 and DEL3589 are always a better match for ChEMBL than DEL2266, no matter what property the emphasis is set on. Both these libraries are roughly equal in terms of their proficiency to mimic the chemical space cover provided by ChEMBL. However, DEL3589 is a better "ChEMBL substitute" specifically in terms of MW distribution, while DEL1847 is more similar with respect to the QED property. Note that any underlying property could be used—such as activity landscapes generated from experimentally validated structure−activity sets, in which case the ChEMBL-likeness of DELs can be specifically tailored with respect to chemical space zones populated by the known actives.

5. CONCLUSIONS

In this study, we propose chemical library space (CLS) visualization, using μGTM. In the context of a large pool of 2497 virtual DELs and ChEMBL database (used as reference), CLS visualization is an intuitive way to gain a global oversight of this portfolio of diverse libraries, unmatched by simple pairwise DEL−DEL or DEL−ChEMBL overlap scoring. This oversight may furthermore be extended to include any other chemical library, combinatorial or not, to be located on existing μ-maps.

Several μGTMs were created, using evolutionary map parameter optimization aimed at preserving interlibrary distances in original high-dimensional CLS on the map. They provided a senseful positioning of libraries relative to the reference collection and each other on the μ-map, matching their similarity observed in the initial CLS defined by either Φ, Λ, or Ω. For the latter, five property μGTMs were created—using MW, log *P*, number of H-bond acceptors, number of H-bond donors, and QED-modulated Ω as library descriptors.

μGTMs support landscapes of library characteristics. These are typically class landscapes displaying the relative population density of libraries of a particular "class" (having a given feature) versus all of the other libraries. Herein, exemplified DEL classifications concern the number of chemistry cycles (2BB versus 3BB DELs), reaction types (coupling or heterocyclization reactions)—but any other deemed relevant by the practitioner (e.g., high cost versus low cost) are straightforward to implement. μGTM colored by the number of reaction cycles showed that there are many CLS zones that are 3BB DEL-specific, accentuating the necessity to use 3BB designs to cover vaster regions of CLS studied herein. The use of a reaction-type-colored μGTM revealed that although including a few classical coupling reactions in DEL synthesis

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

can cover a substantial portion of the produced DEL space, the incorporation of heterocyclization reactions is necessary to explore the otherwise uncharted territories of the CLS. This emphasizes the necessity for the development of BBs for heterocyclization reactions and the expansion of the functional group transformation reaction toolkit that will allow us to access new BB classes, probably more populated, for this type of transformation.

The proposed method of CLS analysis and visualization using μGTM represents an efficient and useful tool for (1) providing a senseful bird's eye view of the whole CLS and simplifying the analysis of interlibrary relationships; (2) analysis of the CLS from different perspectives, positioning libraries by either chemical space similarity or/and property distribution similarity; and (3) the selection of a compound library covering the desired chemical and property space out of thousands of possible ones, given the appropriate reference database. In this regard, it can be beneficial to add this method to the toolkit of medicinal chemists that deal with the selection of a screening collection to verify beforehand if the designed library covers the desired chemical and property space or identify closely lying analogous libraries with similar properties on μGTM.

■ ASSOCIATED CONTENT

Data Availability Statement

The data used in this work are available in the public domain resources: biologically relevant compounds from ChEMBL[17] (version 28), https://www.ebi.ac.uk/chembl/, eMolecules[26] building blocks that were used for DEL generation using eDesigner[22] are partially available on the website https://www.emolecules.com/products/building-blocks, and Enamine[27] building blocks are available on the website https://enamine.net/building-blocks. GTM software used for CS and CLS visualization herein is available under paid license.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00719.

> Density landscapes of 2BB DEL222, ChEMBL, and 3BB DEL3703 showing the absence of connection between the extent of CS coverage and number of BBs; list of reactions included in reaction types used to color the μGTM; operational parameters used to build all maps; list of μGTM parameters that were optimized by GA; and equations used to calculate all CLS vectors (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.3c00719

Funding

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank eMolecules Inc.[26] for the provided collection of commercially available building blocks that were used for the generation of DELs analyzed in this work.

■ GLOSSARY

CS, chemical space, a conceptual space featuring molecules as objects; **CLS**, chemical library space, a conceptual space featuring chemical libraries as objects; **GTM**, generative topographic mapping, a dimensionality reduction method used for CS analysis and visualization; **μGTM**, meta-GTM, a dimensionality reduction method used for CLS analysis and visualization; **CRV**, cumulated responsibility vector, used for representing a compound library based on its compound responsibility distribution on a GTM; **Φ**, normalized CRV, used for unified library representation irrespective of their size; **Λ**, library-modulated CRV, used for representing libraries with respect to their overlap with a reference collection; **Ω**, property-modulated CRV, used for representing libraries considering the properties of underlying compounds

■ REFERENCES

(1) Ottl, J.; Leder, L.; Schaefer, J. V.; Dumelin, C. E. Encoded Library Technologies as Integrated Lead Finding Platforms for Drug Discovery. *Molecules* **2019**, *24*, No. 1629.

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

(3) McCarthy, K. A.; Franklin, G. J.; Lancia, D. R., Jr; Olbrot, M.; Pardo, E.; O'Connell, J. C.; Kollmann, C. S. The impact of variable selection coverage on detection of ligands from a DNA-encoded library screen. *Slas Discovery* **2020**, *25*, 515−522.

(4) Goodnow, R. A., Jr *A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery.*; John Wiley & Sons: Hoboken, New Jersey, 2014.

(5) Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *J. Chem. Inf. Model.* **2023**, *63*, 4042−4055.

(6) Jolliffe, I. T. *Principal Component Analysis for Special Types of Data*; Springer, 2002.

(7) Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59−69.

(8) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215−1221.

(9) van der Maaten, L.; Hinton, G. Viualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(10) Bishop, C. M.; Svensen, M.; Williams, C. K. I. GTM: The generative topographic mapping. *Neural Comput.* **1998**, *10*, 215−234.

(11) Horvath, D.; Marcou, G.; Varnek, A. Generative topographic mapping in drug design. *Drug Discovery Today: Technol.* **2019**, *32-33*, 99−107.

(12) Gaspar, H. A.; Baskin, I. I.; Varnek, A. Visualization of a Multidimensional Descriptor space. In *Frontiers in Molecular Design and Chemical Information Science-Herman Skolnik Award Symposium 2015: Jürgen Bajorath*; ACS Publications, 2016; pp 243−267.
(13) 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.
(14) Maggiora, G. M.; Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 795−802.
(15) 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*, No. 32.
(16) 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.
(17) 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.
(18) 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*, No. 2100289.
(19) *ChemaAxon. JChem*, version 20.8.3; ChemAxon Ltd: Budapest, Hungary, 2020.
(20) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855−868.
(21) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual screening with generative topographic maps: how many maps are required? *J. Chem. Inf. Model.* **2019**, *59*, 564−572.
(22) Martín, A.; Nicolaou, C. A.; Toledo, M. A. Navigating the DNA encoded libraries chemical space. *Commun. Chem.* **2020**, *3*, No. 127.
(23) Horvath, D.; Brown, J.; Marcou, G.; Varnek, A. An evolutionary optimizer of libsvm models. *Challenges* **2014**, *5*, 450−472.
(24) Huang, Y.; Savych, O.; Moroz, Y.; Chen, Y.; Goodnow, R. DNA-encoded library chemistry: amplification of chemical reaction diversity for the exploration of chemical space. *Aldrichimica Acta* **2019**, *52*, 75−87.
(25) Oksiuta, O. V.; Pashenko, A. E.; Smalii, R. V.; Volochnyuk, D. M.; Ryabukhin, S. V. Heterocyclization vs Coupling Reactions: A DNA-Encoded Libraries Case. *J. Org. Pharm. Chem.* **2023**, *21*, 3−19.
(26) eMolecules Inc. https://www.emolecules.com/2020.
(27) Enamine Ltd. https://enamine.net/2020.

https://doi.org/10.1021/acs.jcim.3c00719
J. Chem. Inf. Model. 2023, 63, 5571−5582

Summary

In this study, the meta-GTM approach was used to visualize the CLS formed by a large pool of 2.5K DELs and ChEMBL to gain insights about their interlibrary similarity relationships. The intention was to speed up the process of candidate library selection, for example, here the search for an optimal DEL for primary screening was carried out. This was achieved by gaining a global oversight of the CLS of DELs with respect to ChEMBL using meta-GTM.

Seven meta-GTMs were created to visualize seven CLSs defined either by Φ, Λ, or various Ω vectors. The latter included MW, logP, number of H-bond acceptors, number of H-bond donors, and QED-modulated Ω. All meta-GTMs provided a meaningful positioning of DELs with respect to ChEMBL and each other, proving their high performance in visualizing the CLS. The positioning of libraries on different meta-GTMs logically varied according to the similarity observed in a particular library descriptor space – they were positioned either by structural or property interlibrary similarity. However, the top 100 DELs relative to ChEMBL from the initial CLS always stayed close to it on all the maps, proving their neighborhood behavior compliance and thus usefulness for a quick task-specific library selection based on the similarity to a suitable reference database. In addition, meta-GTM landscapes colored by various library characteristics were created allowing to analyze the CLS from different perspectives. Landscapes colored by the number of chemistry cycles (2BB versus 3BB DELs) and reaction types (coupling or heterocyclization reactions) were created but any other relevant property (e.g. the estimated library production cost, library size, etc.) can be straightforwardly implemented.

Meta-GTMs colored by reaction type used in DEL synthesis allowed to gain new insights about the generated DEL space showcasing their value for multi-library analysis. The majority of the CLS was found to be covered by coupling-based DELs. The latter included almost always more than one of the commonly employed reactions like condensation of carboxylic acid with amine or aldehyde reductive amination. This led to the conclusion that the largest part of the space of 2.5K DELs is purely accessible through popular couplings in DEL chemistry. Nevertheless, a completely new region of DEL CLS

was only reachable with the use of at least one heterocyclization reaction highlighting the importance of this type of transformation to advance to new chemical space zones.

Overall, the proposed method of CLS visualization using GA-optimized meta-GTM proved to be useful for:

1) Providing a general view of the whole CLS represented by thousands of libraries.
2) Fast selection of the candidate library for a specific drug discovery task, allowing identification of the most promising collection out of thousands of possible ones.
3) Muti-perspective view on the CLS by generating meta-GTM landscapes colored according to various library characteristics.

Thus, meta-GTM represents a valuable tool for medicinal chemists working with multiple libraries simultaneously, aiming to select a few based on project-specific criteria. This methodology applies to all libraries, not just DELs. For instance, it can be used to diversify in-house compound collections by comparing them to commercial compound suppliers on the meta-GTM. Another application is identifying a library that is cheaper and easier to synthesize while maintaining the desired structural coverage.

7. BB reactivity prediction and hit prioritization: BRD4 focused DEL study

Introduction

BRD4 is a member of the Bromodomain and Extraterminal (BET) protein family, and plays a critical role in the development of many cancer types, and is associated with tumor metastasis[75,76]. BRD4 recognizes and binds superenhancers that lead to the substantial over-expression of oncogenes causing cancer cell proliferation, survival, tumor initiation and cancer progression[77]. Hence, BRD4 has attracted a lot of interest as a target for drug development. In the past decade, multiple small molecule BRD4 inhibitors and degraders were developed that showed promising anti-cancer effects in pre-clinical models[77]. However, the in-vivo studies on newer BET inhibitors[78–90] often report response data without mentioning associated drug concentrations. This limits the full comprehension of the pharmacokinetic/pharmacodynamic (PKPD) profile of these compounds, thus compromising their promise in human models[25]. Moreover, low solubility after oral dosing decreasing the absorption[91], and short half-lives have been reported for some of the inhibitors[83,92,93]. Hence, there are still research studies focusing on the

Glossary

Competitor compound – Is a known ligand for a biological target, used in the competitive affinity-based screening[4]. It is added at saturated levels to assess its effect on the enrichment of hits from an initial DEL screen. This method is employed when the initial screening identifies too many hits. It helps determine if the hits bind to the expected protein binding site and their competitiveness relative to the known ligand, aiding in the prioritization of hits for further synthesis.

PKPD profile - Pharmacokinetic pharmacodynamic profile. PK analysis allows to understand how the body affects a drug, whereas PD analysis reveals how a drug affects the body[74]. The pharmacokinetics of a compound is determined by its absorption, distribution, metabolism, and excretion in the body. Pharmacodynamics measures the compound's ability to interact with the expected target and exhibit a biological effect[74].

BA – Balanced Accuracy, used as a performance metric for classification ML models, in particular when there is a class imbalance in the dataset. BA is the average of the recall values for both the positive and negative classes. Recall measures the proportion of actual positive or negative samples that are correctly identified by the model.

development of novel BET inhibitors with high potency and a more optimal PKPD profile[25].

DNA-Encoded Library (DEL) technology progress made it an accepted alternative screening platform for early drug discovery projets[94]. DELs allow to screen for affinity millions of molecules all at once, suggesting a more efficient and faster exploration of the chemical space. There are many success stories associated with DEL technology application – a few compounds discovered by DEL screening entered clinical trials according to Gironda-Martínez et al[35].

In a 2019 study by Prinjha[25] and coworkers from GSK, a 117M-sized DEL was screened against BRD4, leading to the discovery of the active and orally available 3,5-dimethylphenol benzimidazole series. Inspired by this work, the Novalix drug development company produced its own in-house DEL based on the findings from the work of GSK scientists. At Novalix, they synthesized a three-BB DEL using 320 protected diamines (BB1), 334 aldehydes (BB2), and 453 carboxylic acids (BB3) via split-and-pool synthesis. These are the BBs that were validated to have sufficient reactivity to be used in this DEL production. The nominal size of this DEL is 18 300 960 compounds, however, only 14.5M compounds were successfully synthesized. This was followed by screening on BRD4. After washing out non-binders and elution of the bound compounds from the protein, the DNA codes of the binders were amplified and sequenced. To identify hits, the enrichment factor (EF) was calculated, it represents the increase in frequency of specific DNA sequences (and their corresponding small molecules) after affinity selection compared to their frequency before affinity screening when no protein was present. Typically, this metric is specific to a pharmaceutical company, but Novalix did not disclose the exact enrichment calculation procedure. Based on the custom EF cut-off of 600, Novalix scientists selected the best binding 70 230 compounds and therefrom selected 102 hits. However, the screening of these 102 hits against the target protein in the presence of the BRD4 competitor compound (having pIC50 of 6.9 with a concentration in the mix of 100 μM) with the aim of tuning the hit recovery resulted in 0 out of 102 hits bound.

The aforementioned experimental results raise the following questions concerning:

1) Building Block (BB) reactivity: How to accurately select reactive BBs without time-consuming and expensive steps of reagent validation?
2) The activity of hits: Do 102 hits or 70K selected binders possess sufficient bioactivity against BRD4? Are these hits structurally similar to the existing BRD4 binders?

To answer the first question, we trained 1530 SVM models on experimentally determined BB reactivity labels from Novalix to classify BBs according to their reactivity. For the second question, we trained a BRD4 bioactivity SVR model on public data from ChEMBL32[95] and predicted the activity (pIC50) of Novalix hits. The best predictive models with good performance on the ChEMBL32 test set were combined in a consensus model that was used for pIC50 prediction for 102 hits. The overlap of Novalix 102 hits with the BRD4 inhibitors from ChEMBL32 in the chemical space was analyzed using Generative Topographic Mapping (GTM) – a robust chemical space visualization method[9,15,47,67,68]. Overall, this work shows new chemoinformatic ways of pre- and post-processing of focused DEL data that can be cost-effective, time-saving, and insightful for medicinal chemists working on DEL production and screening.

Data

BB reactivity

Acquisition of building blocks (BBs) for DEL synthesis is done through careful selection of those with appropriate reactivity. This is usually performed experimentally by validation of a BB in an on-DNA test reaction between the BB (reagent) and an appropriately functionalized DNA conjugate. BB validation is an essential part of DEL design that provides structure-reactivity relationships (SRR) that can further guide the purchase of new reagents[4].

At Novalix, for focused BRD4 DEL synthesis 320 protected diamine (BB1), 334 aldehydes (BB2), and 453 carboxylic acids (BB3) were used. These candidate BBs were assigned labels (valid/moderate/invalid) based on the experimentally determined yield of the reaction. For BB1 the yield was measured for the coupling reaction with the functionalized DNA conjugate (see **Figure 16**). In the case of BB2, the yield of the reaction of different BB2 reagents with a template molecule (a reference DNA-BB1 conjugate) was measured. Likewise, for BB3 the yield was measured with respect to a common reference DNA-BB1-BB2 template.

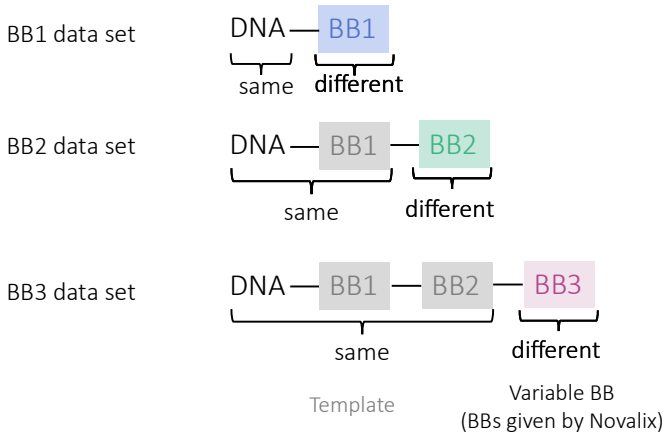


Figure 16. Description of the data given by Novalix. The labels valid/moderate/invalid were given for template+BB pairs, not pairs of individual building blocks. For clarity, here, the compound headpiece and the linker are not depicted.

A building block was considered "valid" when there was more than 70% transformation and thus used for the DEL synthesis. A BB was considered "moderate" when the

validation reaction yield was 50-70% whereas a BB with a yield below 50% was considered "invalid". Moderate and invalid BBs were not used in the final DEL synthesis. The number of labeled BBs is given in **Table 2**.

Table 2. The number of building blocks classified by Novalix scientists as valid/moderate/invalid based on the reaction yield.

	BB1	**BB2**	**BB3**
Total	320	334	453
Valid	179	288	355
Invalid	102	39	98
Moderate	39	7	0

BB structures were provided by Novalix, and their standardization was performed according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics at the University of Strasbourg. This process includes dearomatization and final aromatization (heterocycles like pyridone are not aromatized), dealkalization, conversion to canonical SMILES, removal of salts and mixtures, neutralization of all species, except nitrogen (IV), generation of the major tautomer according to ChemAxon. For the standardized compounds, different ISIDA fragment descriptors were calculated. Descriptor types that resulted in the best models are given in **Table 6**, **Table 7**, and **Table 8**.

BRD4 focused DEL compounds

Using only valid BBs, a 14.5M BRD4-focused DEL was successfully synthesized in three steps of split-and-pool synthesis:

1) Nucleophilic aromatic substitution of DNA-tagged fluoronitrobenzamide with 179 Nboc protected diamines, followed by the reduction of the nitro group.
2) Condensation of diamines with 288 aldehydes.
3) Amide bond formation or acylation of amines with 355 carboxylic acids (preceded by the deprotection of the amine group)

Between each of these chemical reactions, appropriate DNA ligation steps were carried out to encode each of the added BBs. Detailed synthesis steps are depicted in **Figure 17**.

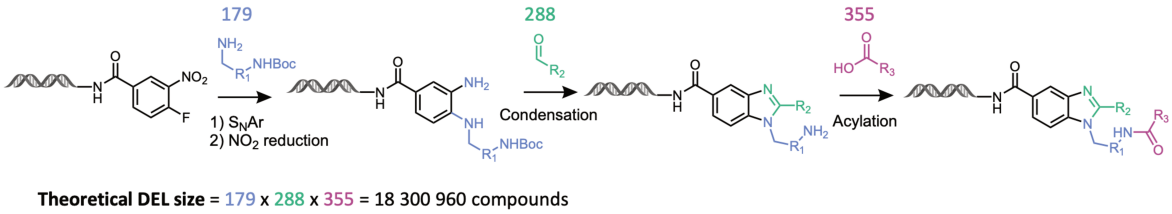


Figure 17. Synthesis steps of Novalix's BRD4 focused DEL. For clarity reasons, DNA ligation steps are omitted.

Chemical structures of the final compounds without DNA tags were provided by Novalix. Their standardization was performed in the same way as for BBs described above. For the standardized compounds, different ISIDA fragment descriptors were calculated. Here we show only those descriptor types that resulted in performant SVR models for pIC50 prediction on ChEMBL32 data (see **Table 3**). The fragment counts were normalized to be in the range from 0 to 1, denoted as "scaled" in **Table 3**, and non-normalized counts are labeled as "orig".

Table 3. Descriptor types that were calculated for DEL compounds with their chemical meaning. The use of these descriptor types resulted in performant SVR models for pIC50 prediction.

Descriptor Type	Meaning
IIRAB—1-3 scaled	Atom-centred fragments with atom symbol and bond inclusion; topological distance 1-3; scaled
IAB—2-6 orig	Sequences with atom symbol and bond inclusion; topological distance 2-6; original

IIAB—1-3 scaled	Atom-centred fragments with atom symbol and bond inclusion; topological distance 1-3; scaled
IAB—FC-AP-2-6 orig	Sequences with atom symbol and bond inclusion; topological distance 2-6; FormalCharge representation; all paths; original

BRD4 inhibitors from ChEMBL32

The efficacy of the majority of molecules from ChEMBL32 was tested using half-maximal inhibitory concentration (IC50), therefore pIC50 was selected as a property to train SVR models on. After standardization, the ChEMBL dataset contained 2831 unique compounds with pIC50(BRD4) in the range of 2.16-9.4. For them, different ISIDA fragment descriptors were calculated but for the sake of clarity, only those descriptor types that resulted in performant SVR models for pIC50 prediction are given in **Table 3**.

To perform a GTM-based comparison of BRD4 inhibitors from ChEMBL32 with 102 hit compounds identified by Novalix, here seven UGTMs[47] were used. For both ChEMBL32 and 102 Novalix hits seven types of ISIDA fragment count descriptors corresponding to each map were thus calculated. ISIDA descriptor types used for each UGTM training as presented in the original article of Casciuc et al.[47] are given in **Table 4**.

Table 4. Descriptor types used for UGTM training along with their meaning and dimensionality of the descriptor spaces.

Map No	Descriptor type	Meaning	Descriptor space dimensionality
1	**IA-FF-FC-AP-2-3**	Sequences of atoms with a length of 2–3 atoms labeled by force field types and formal charge status using all paths.	5161
2	**IIRAB-FF-1-2**	Atom-centered fragments of restricted atom and bonds of 1–2 atoms labeled by force field types	3172

3	**IAB-PH-FC-AP-2-4**	Sequences of atoms and bonds of a length 2–4 atoms labeled by pharmacophoric atom types and formal charges using all paths	4245
4	**IA-2-7**	Sequences of 2–7 atoms.	6520
5	**IAB-FC-AP-FC-2-4**	Sequences of atoms and bonds of 2–4 atoms labeled by formal charge using all paths	3437
6	**IA-FF-P-2-6**	Sequences of atom pairs with a length of 2–6 intercalated bonds labeled by Force Field type	2901
7	**III-PH-3-6**	Atom triplets labeled by pharmacophoric atom types with topological distance from 3 to 6 bonds	4846

Results and discussion

Prediction of reactivity label of BBs

Three SVM models were trained on the data for BB1, BB2, and BB3 provided by Novalix and tested in a 5-fold cross-validation (CV). The data was split into 80% training and 20% validation sets, with the exact number of data points detailed in **Table 5**. For SVM parameter optimization a hill climbing algorithm was used and the objective function was Balanced Accuracy (BA). Both linear and RBF kernel SVMs were used in multiple model creation using different descriptor sets. The types of descriptors that resulted in successful models are listed in **Table 6**, **Table 7**, and **Table 8**. Additional fragmentation options from these tables include DoAllWays (DAW), which searches for all paths connecting two atoms if fragments are sequences, and AtomPairs (AP), which removes all constitutional details of a sequence and only provides the number of constitutive atoms. As a result, 1530 models per each BB were trained, and from these, a few with the highest BA on the validation set were selected.

Table 5. Description of the dataset used for training and validation of SVM models. Only valid and invalid data were included in the dataset.

Dataset	# of data points	# of "valid"	# of "invalid"
BB1	281	179	102
BB2	327	288	39
BB3	453	355	98

BB1 reactivity was easiest to predict (BA=0.9-0.95), followed by (BA=0.85-0.89) for BB2 and (BA=0.7-0.78) for BB3 as can be seen from the model results in **Table 6**, **Table 7**, and **Table 8**. This can be due to many factors. Apparently, BB1 reactivity seems to exclusively depend on the amine nucleophilicity (primary amines are less prone to steric hindrance), so it was easy to learn. Templates already incorporating a reference BB1 or a BB1-BB2 moiety are intrinsically more complex and prone to side reactions than the naked DNA headpiece. It should also be noted that the template-BB validation reaction cannot give the best representation of the yield that will be observed in the real DEL mixture when different BB partners react.

The best models based on BA values (from **Table 6**, **Table 7**, and **Table 8**) were combined in three consensus models for each BB set. Together, they will facilitate the future acquisition of BBs for focused BRD4 DEL synthesis, provided that the fragment-based applicability domain condition for new BBs is verified. This will enable the selection of BBs that are likely to result in high reaction yields, thereby minimizing the purchase and experimental testing of unnecessary reagents. Validation experiments will still be necessary but only for BBs predicted to be reactive, decreasing the number of actual experiments to carry out.

Table 6. Description of SVM models for BB1 reactivity prediction selected to be combined in a consensus model. The description field contains the descriptor type and model information.

No.	Model	Description	BA
1	t10l2u5k2-3	Triplets; min length 2; max length 5; RBF kernel SVM;	0.953

2	t10l3u5k0-3	Triplets; min length 3; max length 5; linear kernel SVM;	0.927
3	t10l2u4k0-3	Triplets; min length 2; max length 4; linear kernel SVM;	0.919
4	t1l2u5k0-3	Sequences of atoms only; min length 2; max length 5; linear kernel SVM;	0.915
5	t4l2u3k0-3	Atom centered fragments based on sequences of atoms; min length 2; max length 3; linear kernel SVM;	0.915
6	t9l2u4APk0-3	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 2; max 4; linear kernel SVM;	0.911
7	t9l3u4APk0-3	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 3; max 4; linear kernel SVM;	0.911
8	t9l2u4APk0-2	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 2; max 4; linear kernel SVM;	0.91
9	t9l2u5APk2-2	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 2; max 5; RBF kernel SVM;	0.91
10	t3l3u4k0-3	Sequences of atoms and bonds; min length 3; max length 4; linear kernel SVM;	0.903
11	t8l2u3k0-3	Atom centered fragments based on sequences of atoms of fixed length; min length 2; max length 3; linear kernel SVM;	0.903
12	t10l2u5k0-3	Triplets; min length 2; max length 5; linear kernel SVM;	0.903
13	t9l3u4APk0-2	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 3; max 4; linear kernel SVM;	0.902
14	t9l2u3APk0-3	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 2; max 3; linear kernel SVM;	0.9
15	t9l2u5APk0-3	Atom centered fragments based on sequences of atoms and bonds of fixed length; min length 2; max 5; linear kernel SVM;	0.9

Table 7. Description of SVM models for BB2 reactivity prediction selected to be combined in a consensus model. The description field contains the descriptor type and model information.

No.	Model	Description	BA
1	t8l2u5k0-5	Atom centered fragments based on sequences of bonds of fixed length; min length 2; max length 5; linear kernel SVM;	0.892
2	t4l3u4APk0-5	Atom centered fragments based on sequences of atoms; min length 3; max length 4; AtomPairs; linear kernel SVM;	0.883
3	t3l3u4k0-5	Sequences of atoms and bonds; min length 3; max length 4; linear kernel SVM;	0.883
4	t8l3u5k0-5	Atom centered fragments based on sequences of bonds of fixed length; min length 3; max length 5; linear kernel SVM;	0.883
5	t5l3u4k0-5	Atom centered fragments based on sequences of bonds; min length 3; max length 4; linear kernel SVM;	0.867
6	t5l4u4APk0-5	Atom centered fragments based on sequences of bonds; min length 4; max length 4; AtomPairs; linear kernel SVM;	0.858
7	t5l4u4k0-5	Atom centered fragments based on sequences of bonds; min length 4; max length 4; linear kernel SVM;	0.858
8	t7l4u4APk0-5	Atom centered fragments based on sequences of atoms of fixed length; min length 4; max length 4; AtomPairs; linear kernel SVM;	0.858
9	t8l4u4k0-5	Atom centered fragments based on sequences of bonds of fixed length; min length 4; max length 4; linear kernel SVM;	0.858

Table 8. Description of SVM models for BB3 reactivity prediction selected to be combined in a consensus model. The description field contains the descriptor type and model information.

No.	Model	Description	BA
1	t7l2u5APk0-3	Atom centered fragments based on sequences of atoms of fixed length; min length 2; max length 5; linear kernel SVM;	0.782
2	t7l2u4APk0-3	Atom centered fragments based on sequences of atoms of fixed length; min length 2; max length 4; linear kernel SVM;	0.764
3	t3l3u5k0-3	Sequences of atoms and bonds; min length 3; max length 5; linear kernel SVM;	0.736
4	t3l2u4k0-3	Sequences of atoms and bonds; min length 2; max length 4; linear kernel SVM;	0.721
5	t8l2u5k0-3	Atom centered fragments based on sequences of bonds of fixed length; min length 2; max length 5; linear kernel SVM;	0.721
6	t3l2u5k0-3	Sequences of atoms and bonds; min length 2; max length 5; linear kernel SVM;	0.714
7	t3l4u4k0-3	Sequences of atoms and bonds; min length 4; max length 4; linear kernel SVM;	0.707
8	t8l3u4k0-3	Atom centered fragments based on sequences of bonds of fixed length; min length 3; max length 4; linear kernel SVM;	0.707
9	t8l3u5k0-3	Atom centered fragments based on sequences of bonds of fixed length; min length 3; max length 5; linear kernel SVM;	0.707
10	t3l4u5DAWk0-3	Sequences of atoms and bonds; min length 4; max length 5; linear kernel SVM;	0.704

Analysis of the chemical space of the BRD4 focused DEL

GTM-based analysis

Structures of 102 hits from Novalix DEL were compared to the space of publicly available compounds tested against BRD4 from ChEMBL32 using GTM. Seven UGTMs were used to project ChEMBL and Novalix compounds and visualize their structure-activity relationships. UGTMs were trained by Casciuc et al.[47] to predict the biological activity of molecules from ChEMBL23. Training and validation sets for these maps consisted of 618 targets from ChEMBL23, including BRD4 protein[47]. Therefore, they are suitable for the visualization and activity prediction of the chemical space of compounds tested against BRD4. **Figure 18** shows seven UGTMs, on which ChEMBL32 compounds tested against BRD4 were projected. The maps were colored according to the measured pIC50 values of ChEMBL23 compounds giving rise to pIC50 landscapes. On each of these UGTM landscapes, 102 hits from Novalix were projected (no activity data is yet associated with these latter). Based on the color of the region of the map where the hits fall, their possible pIC50 value can be inferred. On all maps, the majority of hits represented as black points fall into either orange, yellow, or green zones corresponding to pIC50=6-8. A more detailed analysis of UGTM pIC50 predictions is given in the bar plot in **Figure 19**. The consensus prediction obtained using all maps shows that 86% of 102 hits are predicted as submicromolar inhibitors of BRD4.

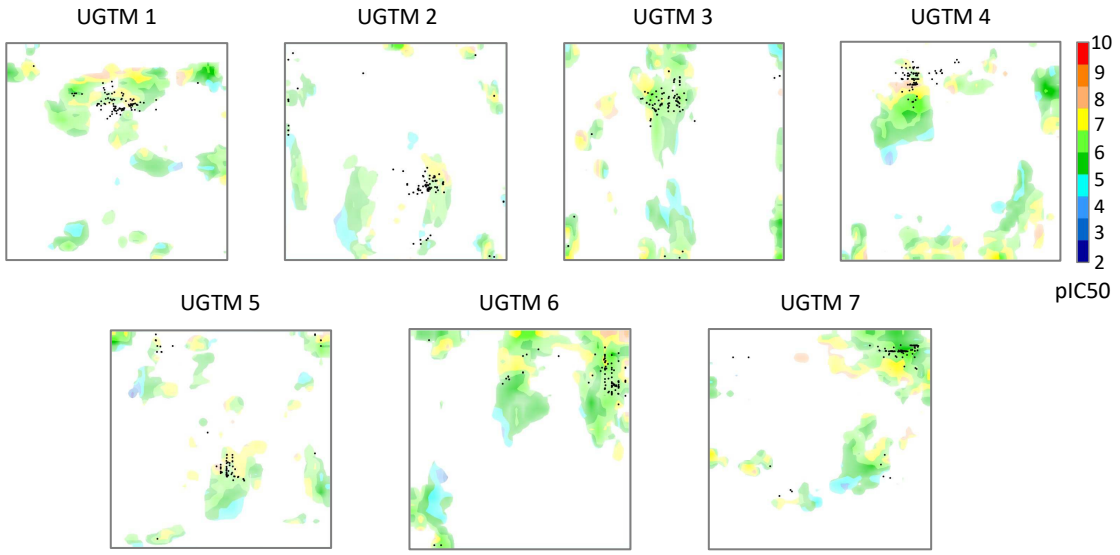


Figure 18. pIC50 landscapes of the ChEMBL32 dataset of compounds tested against the BRD4 target in seven universal maps (UGTMs). On each of the landscapes, 102 Novalix hits are projected as black dots.

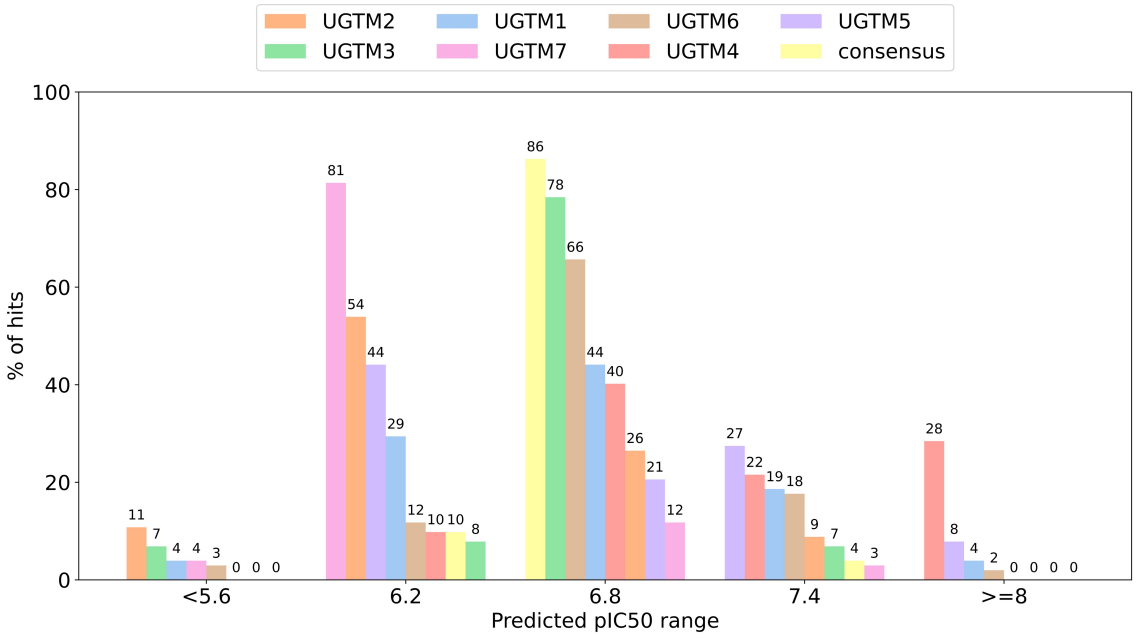


Figure 19. Percentage of 102 Novalix hits predicted to have particular pIC50 values by seven UGTMs.

SVM-based analysis

GTM (a purely neighborhood-based predictor) is interpretable but less predictive than most supervised ML methods used for property prediction[68,96]. Therefore, the pIC50 values for 102 hits were also predicted using RBF kernel SVR trained on the same ChEMBL32 pIC50 data. Models based on different types of ISIDA fragment descriptors were built. They were validated using a 3-fold cross-validation repeated 12 times and the four top-scoring ones on the test set were selected to be combined in a consensus model. Details and performance of the top models are given in **Table 9**. The consensus model was tested for pIC50 prediction on 102 Novalix hits but also on 70K binders, 70K random molecules from the Novalix DEL, and on 70K random ChEMBL32 molecules tested against many different targets. This was done to test whether the model simply memorized the training BRD4 data or truly learned to identify specific characteristics of BRD4 molecular activity.

Table 9. Description of SVM models and their performance on the test set.

Data set	**Descriptor type**	**SVM kernel**	**Performance on test set (Q^2)**
ChEMBL32 pIC50	IIRAB—1-3 scaled	RBF	0.79
	IAB—2-6 orig		0.78
	IIAB—1-3 scaled		0.78
	IAB—FC-AP-2-6 orig		0.77

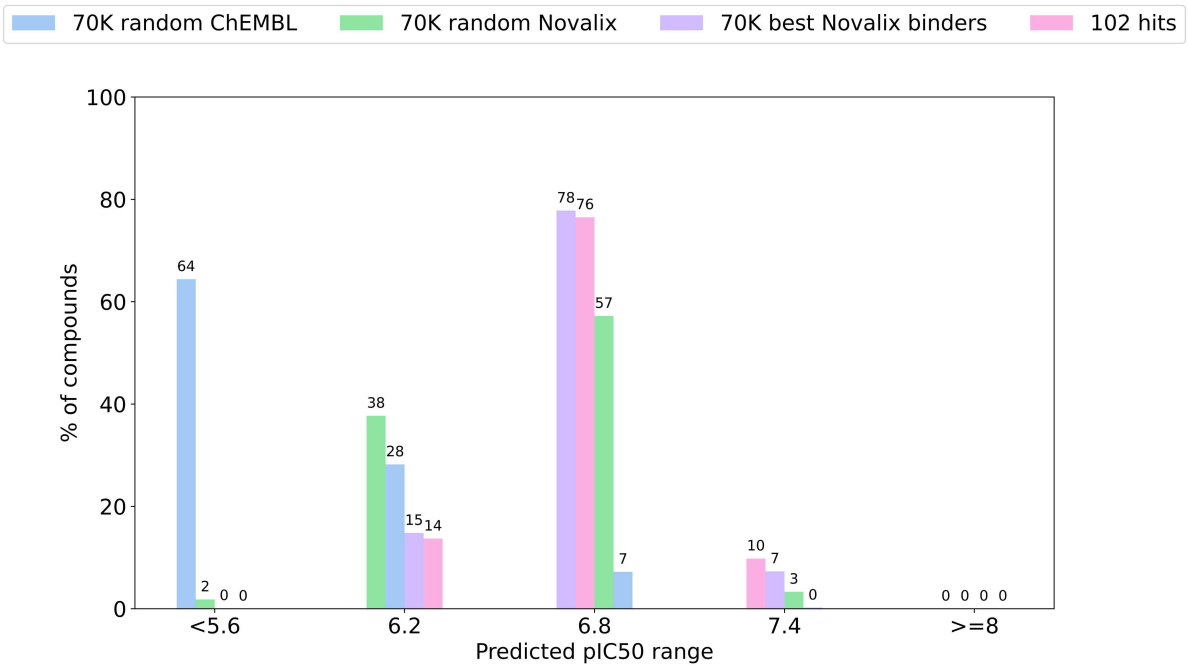


Figure 20. Percentage of compounds from four different data sets predicted to have particular pIC50 values by the SVR model trained on ChEMBL32 BRD4 data.

Figure 20 shows that the consensus SVR model logically predicted 76% of 102 hits and 78% of 70K best binders as submicromolar BRD4 inhibitors. More than half of 70K randomly selected compounds from the DEL were predicted to have submicromolar half maximal inhibitory concentration. 64% Of randomly selected compounds from ChEMBL32 were predicted to be the least potent, showing that the model learned to differentiate BRD4 inhibitor scaffolds from the random target compound structures.

Both SVR and GTM-based predictions indicate that 102 hits are in the same pIC50 range as the competitor molecule (pIC50 of 6.9). The screening with the competitor molecule is usually done to adjust the affinity screening conditions and then see its effect on the enrichment of DEL compounds and thus better characterize them[4]. It is considered in DEL technology that if after such a screen where a known ligand is present in saturated levels (in the case of Novalix - 100 μM), the enrichments of hits are reduced or they are completely prevented from binding to the target, such hits can be considered as "competitive" binders with respect to the known ligand. Compounds that are not prevented from binding are those that interact with another binding site of the target. It is also important to consider that typical concentrations of the target in a DEL screen is ~1

μM and the concentration of DEL members is in the order of ~10-100 fM[4]. In the case of Novalix's 102 hits, failure to bind the BRD4 upon addition of the competitor molecule indicates (1) their affinity for the right binding site and (2) their competitiveness with the known ligand that was also confirmed by ML modeling.

Summary

In this work, BB and hit analysis was performed for the focused DEL targeting BRD4 protein provided by Novalix.

1530 SVM models trained on experimentally determined BB reactivity labels from Novalix were developed. A few of them achieved good performance – Balanced Accuracy (BA) of the best reactivity class prediction models for BB1, BB2, and BB3 was 0.95, 0.89, and 0.78, respectively. The best models for each BB were combined in the consensus model which can potentially accelerate the experimental BB selection and validation procedure at Novalix by filtering out unreactive BBs.

102 Hits obtained after DEL affinity selection against BRD4 protein at Novalix were compared to known BRD4 inhibitors from ChEMBL32 using GTM. This showed that 102 hits overlap with micromolar and submicromolar BRD4 inhibitors in the chemical space defined by structural fragment ISIDA descriptors. In addition, SVR models for BRD4 bioactivity prediction were trained on the same data from ChEMBL32. The best models showing good performance on the test set (BA ≈ 0.8) were combined in a consensus model that was used to predict the pIC50 of Novalix hits. The model logically predicted the majority of 102 hits and 70K binders to be in the submicromolar half-maximal inhibitory concentration range, whereas the randomly selected compounds from ChEMBL and Novalix DEL are expectedly predicted to be less potent. In more detail, 76% of 102 hits were inferred to have a pIC50 of 6.8. These results are coherent with experimental observations during the hit recovery tuning screen in the presence of the competitor ligand performed by Novalix. The lack of binding observed for the 102 hit compounds in an affinity screening against BRD4 in the presence of saturating levels (100 μM) of the competitor molecule (pIC50 = 6.9) indicates the competitiveness of hits with the known BRD4 ligand for the right binding site. This competitiveness is confirmed by ML-predicted pIC50 values for these hits.

Overall, this work shows how supervised (SVM) and unsupervised (GTM) ML methods can be useful for the pre- and postprocessing of focused DEL data. GTM and SVM used in this work for chemical space analysis and reactivity or pIC50 prediction, were shown to rationalize and accelerate BB reactivity validation and DEL hit prioritization tasks.

8. Combinatorial Library Network (CoLiNN) for combinatorial library visualization without compound enumeration

Introduction

DELs are combinatorial libraries consisting of compounds formed by various combinations of BBs according to specific reaction rules. To carry out the computational analysis of DELs, first, their compounds need to be enumerated. However, compound enumeration is a task that consumes significant time, computational resources, and memory. Moreover, not any library can be enumerated, the reported upper limit for enumerable library size is 10^{12} compounds[4]. In the case of DEL space analyzed in this thesis that is composed of 2.5B compounds, enumeration and subsequent standardization and descriptor calculation were time- and memory-limiting steps of the analysis, necessitating long calculations on multiple 48 CPU machines. Hence, in this work, a no-enumeration approach to combinatorial library analysis and visualization was developed.

Glossary

GCN – Graph Convolutional Network, designed to operate on graph-structured data (e.g. molecules). Each node of the graph (atom of the molecule) can have associated features, which are used as input to the network. The convolution operation aggregates the features of each node with those of its neighbors, and multiplies them by learned weight matrix and non-linear activation function[97].

CRV – Cumulated Responsibility Vector that is derived from the GTM of the library. It is a vector that encodes the approximate total number of compounds from each node of the map. It is used as a representation of the library as a whole.

Normalized CRV (Φ) – Library-size independent library descriptor vector. It can also be visualized as a density landscape showing the quantitative distribution of compounds in the chemical space of the library.

Tanimoto coefficient (Tc) – Similarity coefficient usually calculated between two molecular fingerprints or vectors. However, in the context of library comparison, it is used to calculate the similarity in terms of the chemical space coverage of two libraries. For this, Tc is calculated between CRV or Φ vectors of two libraries to compare.

To our best knowledge, there was only one study that reported successful visualization of a 90K combinatorial library using Multi-Dimensional Scaling (MDS) without compound enumeration. In their work, Agrafiotis and Lobanov[70] trained a Multi-Layer Perceptron (MLP) able to predict the coordinates of the products on the MDS map

from descriptors of their BBs. However, in their approach, reaction information was omitted. Nevertheless, reactions, especially heterocyclizations that introduce a whole new scaffold into the molecule can significantly influence the structure of the final product and thus its position on the chemical space map. Hence, in this work, we propose a Combinatorial Library Neural Network (CoLiNN) that allows to predict the projections on the GTM from the BBs and reactions without the need for compound enumeration.

CoLiNN is a Graph Convolutional Network (GCN) trained, validated, and tested on 2.5K DELs generated in our previous studies. The target value to predict is a responsibility vector of a DEL compound, which represents its projection onto the GTM. The 2.5K DELs are 2- or 3-BB libraries with 2 or 3 reactions with the full size from 1M to 7B compounds. Previously, only 1M compounds per DEL were generated and analyzed using GTM to avoid the lengthy and computationally intensive process of enumerating entire compound libraries. Here, we aim to bypass the enumeration step to accelerate the chemical space analysis of DELs and combinatorial libraries in general. Two CoLiNN models were developed:

1) A local model trained on a small subset of compounds from one 80M-sized DEL
2) A general chemistry-sensitive model trained on subsets from 388 DELs

The Kullback-Leibler divergence was selected as the loss function to measure the difference between predicted and "true" responsibility vectors. The local model was tested on an 80M compound DEL to evaluate how well it predicts the chemical space map of the full library using only a subset for training. The general CoLiNN model was tested on 2 089 DELs that were not part of the training set but shared building blocks (BBs) and reactions with the 388 DELs used for training. Additionally, it was tested on the remaining compounds from the 388 DELs that did not participate in the training.

In this work, only 2 473 out of the initially generated 2 497 DELs were analyzed. The remaining 24 DELs were excluded due to data quality issues, as these libraries were incorrectly enumerated by eDesigner with one missing BB per compound. Although the compounds were still valid, they did not correspond to the expected ones.

CoLiNN: A Tool for Fast Chemical Space Visualization of DNA-Encoded Libraries Without Enumeration

*Regina Pikalyova[1], Tagir Akhmetshin[1], Dragos Horvath[1], Alexandre Varnek[1]**

Abstract:

Visualization of the combinatorial library chemical space provides a comprehensive overview of available compound classes, their diversity, and physicochemical property distribution - key factors in drug discovery. Typically, this visualization requires time- and resource-consuming compound enumeration, standardization, descriptor calculation, and dimensionality reduction. In this study, we present the Combinatorial Library Neural Network (CoLiNN) designed to predict the projection of compounds on a 2D chemical space map using only their building blocks and reaction information, thus eliminating the need for compound enumeration. Trained on 2.5K virtual DNA-Encoded Libraries (DELs), CoLiNN demonstrated high predictive performance, accurately predicting the compound position on Generative Topographic Maps (GTMs). GTMs predicted by CoLiNN were found very similar to the maps built for enumerated structures. In the library comparison task, we compared the GTMs of DELs and the ChEMBL database. The similarity-based DELs / ChEMBL rankings obtained with "true" and CoLiNN predicted GTMs were consistent. Therefore, CoLiNN has the potential to become the go-to tool for combinatorial compound library design – it can explore the library design space more efficiently by skipping the compound enumeration.

Keywords: Combinatorial library, compound enumeration, DNA-encoded libraries (DELs), GTM

INTRODUCTION

Combinatorial chemistry allows to produce a vast number of structurally diverse molecules by simple and repetitive steps of covalent building block (BB) linkage. DNA-Encoded Library (DEL) is an example of a combinatorial library (CL) that has become a complementary hit identification approach to conventional High-Throughput Screening (HTS). DEL compounds are molecules covalently attached to DNA tags encoding information about the building blocks (BBs) composing the molecules. DNA encoding allows to screen all compounds in a DEL simultaneously in a mixture against a biological target of interest, thus allowing to explore large regions of the chemical space all at once[1].

Various methods, such as QSAR modelling[2], docking[2], and dimensionality reduction[3–5], can be used to analyze combinatorial libraries. However, despite their utility, these methods require an enumeration of compounds. Currently, a library size of about 10^{12} compounds represents a practical upper limit that can be enumerated[6]. When dealing with larger libraries or several large-sized collections, enumeration no longer seems feasible due to the immense processing resources required and 'big data' storage issues[6].

1. *University of Strasbourg, Laboratoire de Chemoinformatique, 4, rue B. Pascal, Strasbourg 67081 (France) *e-mail: varnek@unistra.fr*

The analysis of vast combinatorial libraries without enumeration was first addressed in the works of Rarey et al.[7–11] The authors exploited the property that CLs are composed of different building blocks (BBs) combined according to specific reaction rules to efficiently analyze, search, and compare even non-enumerable combinatorial spaces. The main principle of these methods is to represent each fragment of a molecule by fingerprint representations that capture their chemical and topological features or encode a compound as a tree of its fragment features describing the ability of forming interactions. For example, the Ftrees[7] methodology enables fast query-based similarity searches in CL spaces based on feature trees. Tools like SpaceProp[11] and SpaceCompare[10] extend this approach: SpaceProp calculates property distributions without full enumeration, while SpaceCompare facilitates the comparison of different combinatorial chemical spaces using fragment fingerprint representations. Nevertheless, the visualization of the chemical space using dimensionality reduction methods is unattainable without first enumerating the compounds.

The first method for visualisation of the chemical space without the need for compound enumeration was described in the work of Agrafiotis and Lobanov[12]. They developed a three-layer fully connected Multi-Layer Perceptron (MLP) trained to predict multidimensional scaling projections on a 2D map of combinatorial products using as input descriptors of their respective building blocks. This approach was tested on a two-building block combinatorial library based on reductive amination reaction with 90K compounds. However, their method did not consider the reactions used to create the library. This shortcoming complicates the analysis of combinatorial libraries with multiple reaction types, requiring multiple reaction-specific models instead of a single unified model.

Thus, herein, we propose a Combinatorial Library Neural Network (CoLiNN) that given the building blocks and reactions of the combinatorial library predicts the fuzzy projection of a product on the Generative Topographic Map (GTM) without compound enumeration. This approach can be extended to any kind of chemical space map created using different dimensionality reduction methods. CoLiNN was trained and validated on combinatorial DNA-Encoded compound Libraries (DELs) generated in our previous study[3]. The target value to predict is a responsibility (projection) vector of a compound on the GTM. The DELs used in training consisted of 2- or 3-BB libraries with 2 or 3 reactions, representing large libraries of 1M to 7B compounds. Previously, due to the time-consuming nature of compound enumeration, only 1M compounds per DEL were visualized. This work aimed to train a CoLiNN model capable of predicting GTMs for combinatorial libraries of any size, thereby enabling the visualization of ultra-large DELs. Two CoLiNN models were developed:

1) A local (library-specific) model trained on a small subset of compounds from one 80M-sized DEL
2) A general chemistry-sensitive model trained on subsets from 388 DELs

Both CoLiNN models achieved high predictive performance of GTM projection vectors based on building blocks and reactions. This was measured by the Kullback-Leibler divergence between predicted and true projection vectors for compounds in the validation set. For 80% of 2.5K DELs investigated in this study, general CoLiNN allowed to accurately predict compound responsibility vectors. This was reflected by high Tanimoto similarity values between the predicted and true maps for these DELs - from 0.8 to 0.99. Hence, CoLiNN holds the potential to become an essential tool for combinatorial compound library design, as it can efficiently explore the library design space without requiring compound enumeration. This can be especially advantageous for DEL technology, where medicinal chemists must select the most promising design from thousands of possibilities. With CoLiNN, different sets of building blocks can be tested simultaneously, providing instant visualization of the chemical

space to identify the best options for DEL design.

DATA

Building blocks (BBs)

The set of commercially available BBs from eMolecules Inc[13,14]. and Enamine[14] used for DEL enumeration in our previous study[3] was taken as input to the CoLiNN model. BBs were standardized using ChemAxon Standardizer according to the procedure implemented on the Virtual Screening Web Server of the Laboratory of Chemoinformatics at the University of Strasbourg. This process includes dearomatization and final aromatization (heterocycles like pyridone are not aromatized), dealkalization, conversion to canonical SMILES, removal of salts and mixtures, neutralization of all species, except nitrogen(IV), generation of the major tautomer according to ChemAxon. After standardization and duplicate removal 70 691 unique BBs were obtained. For each of them, a unique identifier starting from 0 was given. For training of the global CoLiNN model, only 388 DELs with unique reaction schemes were used that employ 64 869 BBs for their enumeration from the total BB set.

DELs

One of the most popular combinatorial library technologies is DNA-Encoded Library[15] (DEL) screening. They are created through combinatorial split-and-pool synthesis and the identity of the BBs used in each compound is recorded in a DNA "barcode" that is produced through the ligation of DNA tags after each synthesis cycle. All compounds of the library are then concurrently screened for affinity against a biological target. Successful binders are identified by sequencing of corresponding DNA "barcodes". However, it is more cost-effective to design and analyse virtual combinatorial libraries before their actual synthesis and testing. This approach allows for the identification of potentially hit-enriched compound libraries before any experimental procedures are undertaken[1].

DELs were designed and enumerated in our previous study[3]. In this work, we used 2473 DELs, each containing 1M compounds (randomly enumerated representative subset). The standardization of enumerated molecules was done using the same procedure described above for BBs.

DEL for local CoLiNN model

The local CoLiNN model was trained on compounds from DEL2568. It is a 3-BB DEL based on three reactions: aldehyde reductive amination, Migita thioether synthesis, and guanidinylation of amines. This DEL was fully enumerated by eDesigner and after removing duplicates had a size of 81M compounds. For training, a random subset of 1M compounds was used. Four CoLiNN models were trained on four training sets to select the minimum required training set size: 1) 1M set, 2) 50K random compounds taken from 1M, 3) 25K random compounds taken from 1M, 4) 10K random compounds taken from 1M. The remaining 80M compounds from this DEL (excluding the 1M representative set) were used for testing.

DELs for global CoLiNN model

For global CoLiNN training, we identified how many DELs are needed to get a diverse training set out of the space of 2473 DELs we generated originally. For this, all DELs were clustered according to their reaction schemes. Here reaction scheme stands for a particular sequence of two or three reactions in a defined order used for DEL enumeration. This resulted in 388 different clusters, i.e. unique reaction schemes not accounting for deprotection reactions. See the bar plot showing the number of DELs per reaction scheme in **Figures S1** and **S2** of the **Supporting Information (SI)**. Per each reaction scheme, one DEL was taken for training CoLiNN. Only 10K compounds from each of these DELs were taken for training, giving rise to a 3 880 000 training set compounds.

ChEMBL

ChEMBL database version 28 was taken here as a reference dataset to be used for

comparison to DELs. ChEMBL28 was filtered according to the rules of DEL-likeness introduced in our previous work[4]. Standardization of ChEMBL molecules was done in the same way as for BBs described above. After duplicate removal, the size of the filtered ChEMBL was 1 605 370 compounds.

Reactions

Reaction information was input for CoLiNN training. They were taken from the Supporting Information document of the eDesigner[16] enumeration tool. It contains numerical codes and names of reactions as encoded in eDesigner. In this study, we assigned each reaction a unique index ranging from 1 to 29 to be used for CoLiNN training. The reaction names and the correspondence between eDesigner numerical codes and indices used herein are given in **Table S1** of the **SI**.

METHODS

Generative Topographic Mapping (GTM)

Generative Topographic Mapping[17] (GTM) is an unsupervised method for dimensionality reduction based on manifold learning. The GTM algorithm consists in optimizing the manifold shape to fit the data represented in the multidimensional descriptor space. The optimization is done by tuning manifold hyperparameters, such as its flexibility and smoothness (determined by the number of Radial Basis Functions (RBFs) usually radially symmetric Gaussians, RBF width factor σ, and the spacing of RBFs), map size, and regularization coefficient. When the optimal form of the manifold is found, data points are projected to it with node-specific probabilities called responsibilities (r_{ik} – the responsibility of the molecule i to be projected to the node k). In such a way, a data point (molecule) can be projected to multiple nodes at the same time meaning it is associated with several chemotypes[18,19]. Finally, a manifold is flattened back to its planar form giving an interpretable 2D map with projected on it input dataset compounds.

When multiple molecules of the chemical library are projected onto the GTM, a cumulative responsibility for each node k, see Equation (1) can be calculated, which is roughly equal to the number of compounds residing there. The cumulative responsibility values per node can be rendered using a colour code allowing to visualize the quantitative distribution of compounds across the chemical space giving rise to a density landscape. In this work, density landscapes are used to visualize the chemical space of the analysed DELs.

$$\boldsymbol{c_k} = \sum_{i}^{N} \boldsymbol{r_{ik}} \qquad (1)$$

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

However, to capture library size-independent chemical space coverage, $\boldsymbol{c_k}$ should be normalized by the library size N, resulting in $\phi = (\phi_1, \ldots, \phi_k)$ vector to represent the entire chemical library, see Equation (2).

$$\boldsymbol{\phi_k} = \frac{\boldsymbol{c_k}}{\boldsymbol{N}} \qquad (2)$$

Here, ϕ is calculated either from the predicted or GTM-derived responsibility vectors. To compare the GTM-derived $\boldsymbol{\phi}$ and $\boldsymbol{\phi'}$ predicted by CoLiNN of the same library a Tanimoto similarity coefficient (Tc) can be calculated, see Equation (3). Its value indicates how much the predicted and the "true" GTM-derived maps are similar to each other.

$$\boldsymbol{Tc}(\boldsymbol{\phi}, \boldsymbol{\phi'}) = \frac{\sum_k^K \boldsymbol{\phi_k} \cdot \boldsymbol{\phi'_k}}{\sum_k^K \boldsymbol{\phi_k}^2 + \sum_k^K \boldsymbol{\phi'_k}^2 - \sum_k^K \boldsymbol{\phi_k} \cdot \boldsymbol{\phi'_k}} \qquad (3)$$

K – total number of nodes on the map

Combinatorial Library Neural Network

Inputs

A product molecule (DEL compound) was represented as a sequence of reactions and BB identifiers as shown in **Figure 1**.

In this work the BB was represented as an undirected hydrogen-labelled graph[20] (HLG) where edge type or bond order (single, double, triple, aromatic) is taken into account using the number of hydrogens each atom is connected to. In more detail, the HLG feature vector is defined by atomic number, period, group, number of electrons plus atom's charge, shell, an indication of whether an atom is in a ring, number of neighbouring atoms, and the total number of hydrogens connected to each atom. The latter allows to account for bond order directly in the feature vector avoiding the use of relational GCN where each layer has r weight matrices, where r is the number of unique bond types. Instead, in our GCN, only one weight matrix per layer is used owing to the additional value in the feature vector. The features described herein were calculated using the EPAM Indigo toolkit[21] and graphs were created using PyTorch Geometric[22]. The target value used for training CoLiNN is a responsibility vector of the compound on the GTM as shown in **Figure 1**.

Architecture

CoLiNN is a Graph Convolutional Network that performs two tasks during training: (1) it creates and saves embeddings for Building Blocks (BBs), and (2) it creates and trains on molecule embeddings that are assembled from BB and reaction embedding vectors to be able to predict responsibility vectors of compounds (see **Figure 2** and **Figure 3**). The steps of this architecture are as follows:

1. *Embedding Calculation for Building Blocks (BBs):* Each BB is assigned a unique ID, and its embedding vector is calculated as shown in **Figure 2**. First, atom feature vectors are linearly transformed into initial embedding vectors of dimension D. Then the latter as well as the adjacency matrix are passed through five Graph Convolution Network (GCN) layers.

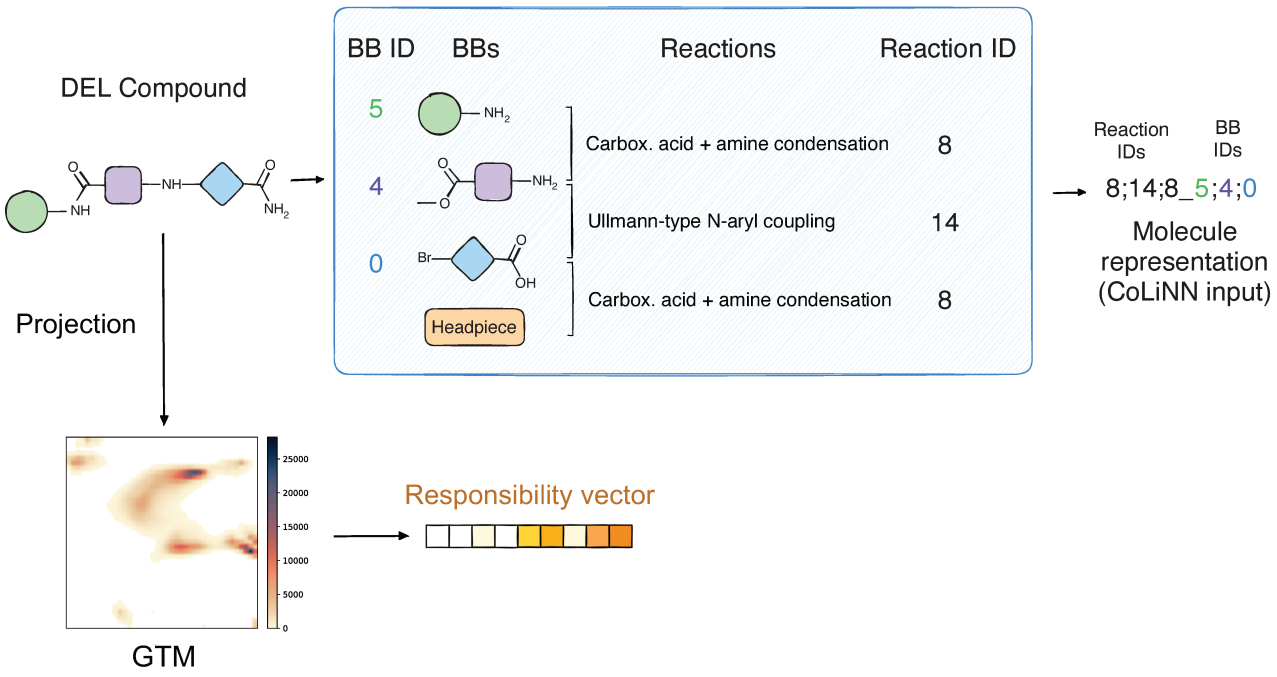


Figure 1. Creation of the product molecule representation that is used as input for CoLiNN.

The mathematical transformation occurring in each GCN layer is described in Equation (4). The obtained atom embedding vectors in this way are then summed up to get a BB embedding vector and saved in a file.

2. *Molecule representation:* The molecule is represented by a sequence of reaction and BB indices. Then, saved BB embeddings are associated with corresponding indices. Reaction embedding vectors are obtained by converting the reaction indices into numerical vectors of dimension D.
3. *Application of Masks:* A mask is applied to the BB embeddings to zero out those generated for padding BBs. For instance, in a 2-BB compound, the embedding for the third "padding" BB is zeroed out.
4. *Assembly of the Full Molecule Vector:* The full molecule vector is created by combining BB and reaction embeddings. They are concatenated to integrate information about both molecular composition and reactions used to enumerate the compound. They are then summed up to form a vector representing the molecule.
5. *Responsibility vector prediction:* The molecule vector is linearly transformed and a softmax activation function is applied to give an output 1681-dimensional responsibility vector, which represents the compound projection on the GTM of size 41x41 nodes.
6. *Loss function calculation:* Kullback-Leibler divergence is calculated between the responsibility vector predicted by CoLiNN and the actual enumerated product projection; see Equation (5).

By following these steps, CoLiNN effectively encodes the complex structure and reaction information of molecules into a form suitable for responsibility prediction tasks without compound enumeration.

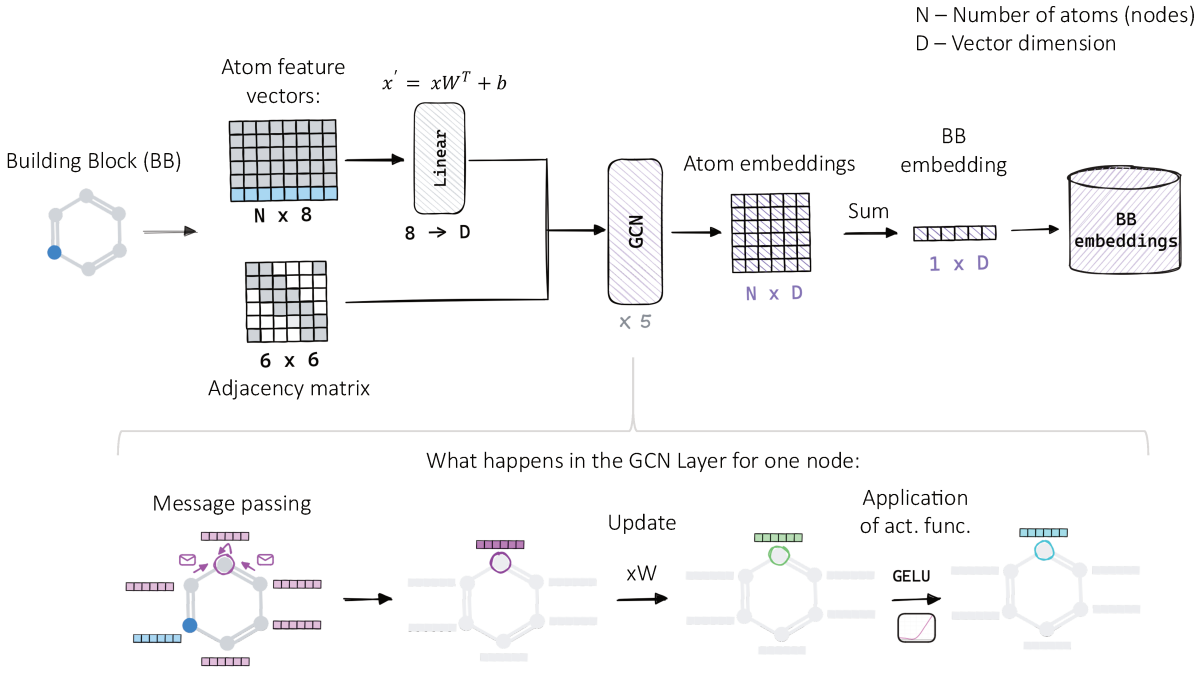


Figure 2. Calculation of building block (BB) embedding vectors.

CoLiNN is coupled with 5 Graph Convolution Network layers where GELU[23] (Gaussian Linear Unit) is used as an activation function. Instead of ReLU used in the original implementation of GCN by Kipf and Welling[24] the GELU activation was chosen as it is smoother than ReLU and is differentiable at every point[25]. This helps to have an improved gradient flow during backpropagation[25,26] and to reduce the number of dead neurons[25]. Mathematically, GCN operation can be described by the following equation:

$$\boldsymbol{H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}(A + I)\tilde{D}^{-\frac{1}{2}}H^{l}W^{l})} \quad (4)$$

$A \in \mathbb{R}^{n\times n}$ – adjacency matrix,

$A_{i,j} = \begin{cases} 1, \text{if there is an edge between nodes i and j} \\ 0, \text{otherwise} \end{cases}$

$I \in \mathbb{R}^{n\times n}$ – identity matrix

$\tilde{D} \in \mathbb{R}^{n\times n}$ – diagonal degree matrix of $\tilde{A}$

$H^{l} \in \mathbb{R}^{n\times d}$ – per-node feature vectors or node embeddings from the previous layer

$W^{l} \in \mathbb{R}^{d\times w}$ – weight matrix for layer l

σ – non-linear activation function, here Gaussian Error Linear Unit (GELU)

n – The number of nodes

d – The number of node features

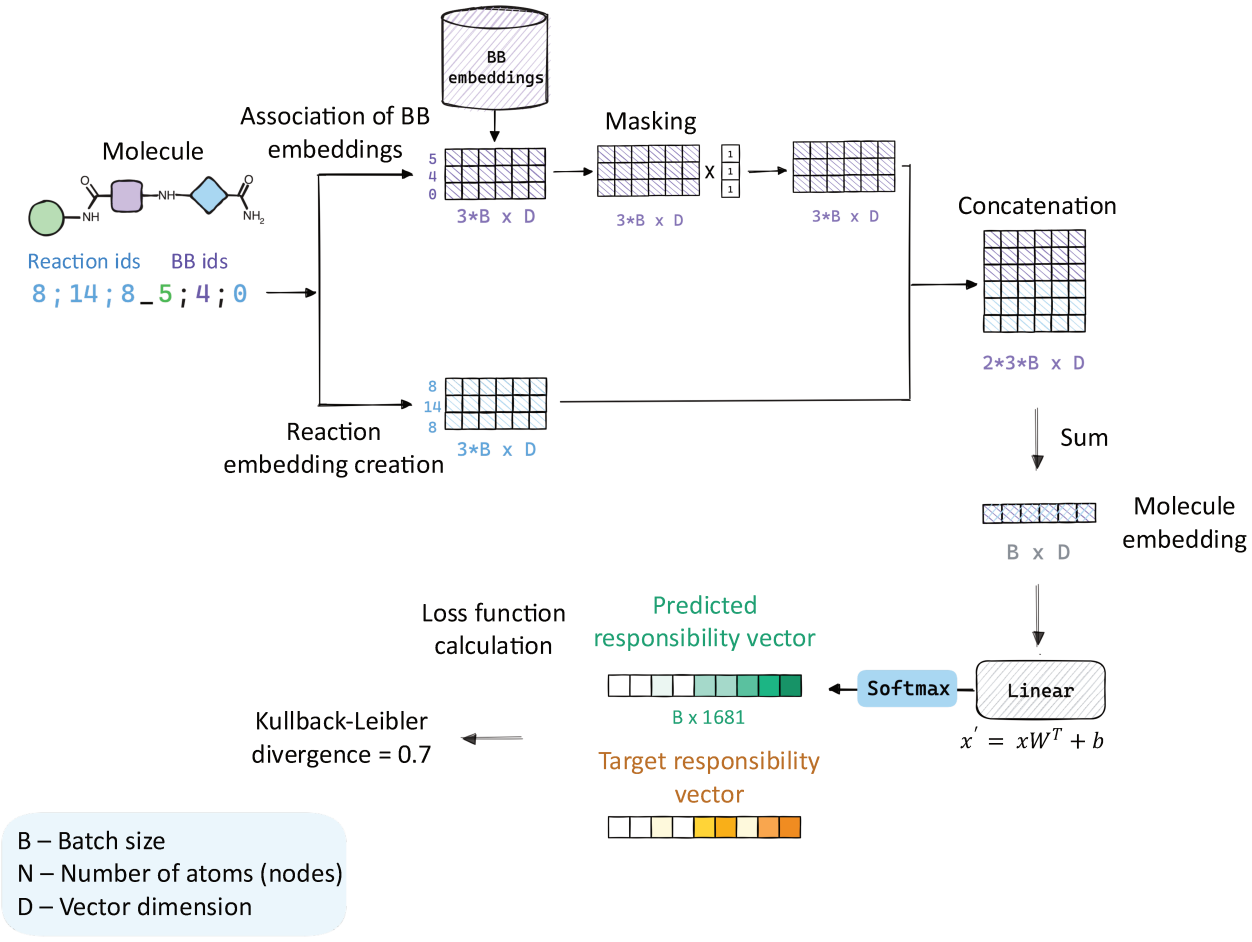


Figure 3. Scheme explaining the training process of CoLiNN. Here, for simplicity, every step is shown for the batch size of one molecule.

Loss Function

The responsibility vector predicted by CoLiNN was compared to the one obtained using the GTM algorithm using Kullback-Leibler divergence, see Equation (5). The latter measures how the inferred probability distribution diverges from the expected probability distribution.

$$KL\,div\,(R_i \parallel R'_i) = \sum_{k=1}^{K} r_{ik} \cdot ln\,\frac{r_{ik}}{r'_{ik}} \quad (5)$$

$R_i = (r_{i1}, r_{i2}, \ldots, r_{ik})$ true responsibility vector of the molecule i

$R'_i = (r'_{i1}, r'_{i2}, \ldots, r'_{ik})$ predicted responsibility vector of the molecule i

RESULTS AND DISCUSSION

Local CoLiNN model

The local CoLiNN model is specific to DEL2568. To select the minimum required training set size, four CoLiNN models were trained and tested using four subsets of sizes 10K, 25K, 50K and 1M DEL2568. This 1M set was in all cases excluded from testing, performed over the remaining 80M compounds from the full DEL. The training and validation KL div loss values for all local models are given in **Table S2** in **SI**.

In **Figure 4**, the reference density landscape of the 80M DEL generated using the GTM algorithm and those predicted by CoLiNN are shown. As the training set size increases from 10K to 1M, the similarity to the reference GTM-produced landscape also increases. However, if instead of comparing the maps visually, we look at the Tanimoto similarity coefficient between the GTM-generated maps and predicted maps Tc(ϕ, ϕ') in **Figure 5**, a plateau of Tc is observed at 25K already. This means that the minimum size required to train a robust model is 25K compounds. This CoLiNN model is able to predict the characteristic GTM projection pattern of the 80M-sized library in 2 hours instead of the 13 days taken by the standard procedure: eDesigner[16] enumeration, standardization, ISIDA fragment descriptor calculation, projection of compounds on the GTM.

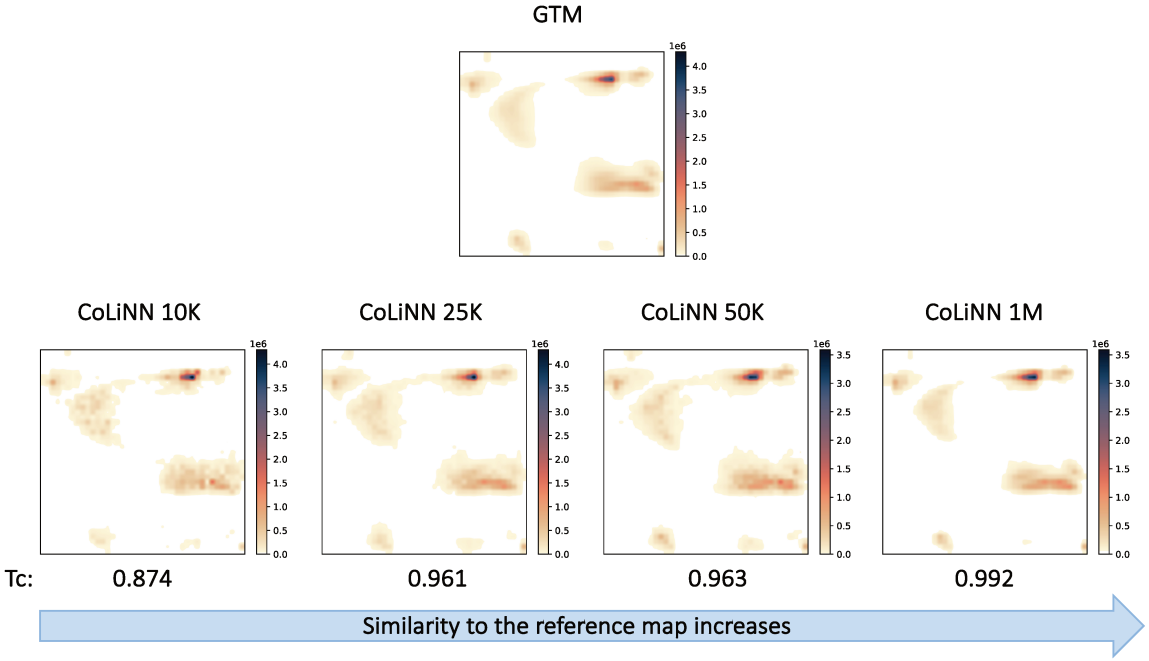


Figure 4. Density landscapes of DEL2568 containing 80M compounds, from left to right: landscape generated by GTM algorithm, predicted by CoLiNN trained on 10K, 25K, 50K, and 1M compounds.

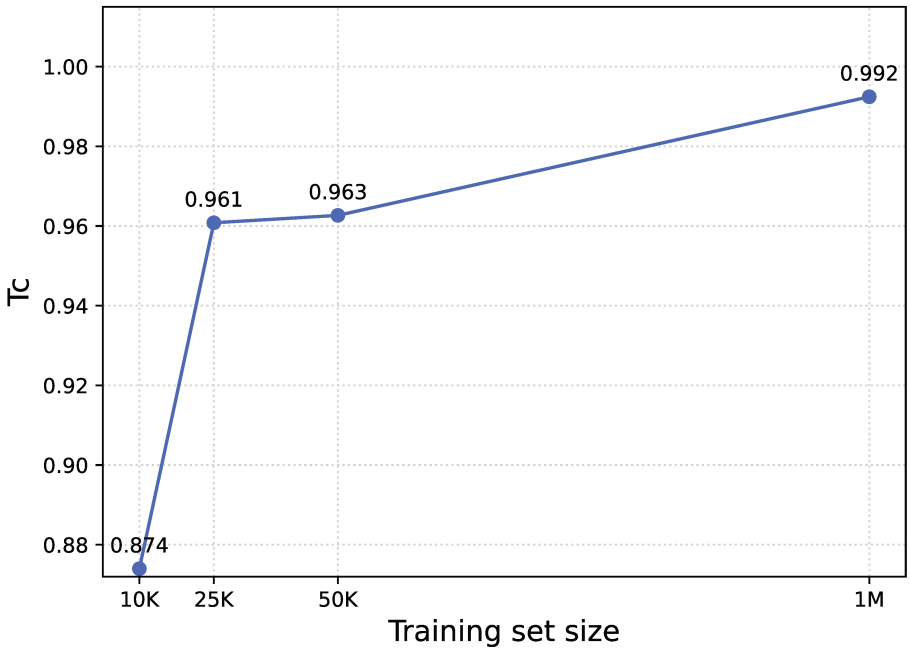


Figure 5. Line plot showing the Tc(ϕ, ϕ') for different CoLiNN models trained on 10K, 25K, 50K, and 1M compounds.

Global CoLiNN model

The global model was trained on 10K samples from 388 DELs employing unique reaction schemes (see **Table S1** in **SI**). 50 epochs were sufficient to achieve convergence, which corresponds to a training duration of CoLiNN of about 13 h. Training and validation KL div loss values for different epochs are reported in **Table 1**.

Table 1. Values of the loss (KL divergence) at the beginning, halfway, and end of training CoLiNN on 388 DELs.

	Train loss	Validation loss
Epoch 1	4.66	2.49
Epoch 50	0.78	0.79
Epoch 100	0.73	0.77

The trained global CoLiNN was tested both on all compounds from 388 DELs not used for training (990K compounds per DEL) and all the other 2089 DELs that did not participate in training at all. Albeit they did not serve for training, these latter share building blocks and reaction schemes with the 388 training DELs. **Figure 6** shows the distribution of Tanimoto similarity values between the "true" and predicted ϕ vectors for all DELs. In the vast majority of cases, predictions are accurate (for almost all train set and for 1600 of test set DELs ≈80% of all DELs) but a few notable failures occur. Examples of badly and perfectly predicted DEL maps (both from the test set) are given at the bottom of **Figure 6**. CoLiNN almost perfectly predicted the responsibility vectors for all compounds from 3BB DEL1953 with Tc(ϕ, ϕ') being 0.99. This is a 3BB library based on Schotten-Baumann coupling between amines and sulfonyl chlorides, 1,2,3-triazole synthesis and aldehyde reductive amination. For the DEL117 the predicted map is significantly dissimilar from the true one as reflected by the low Tc(ϕ, ϕ') of 0.07 – a few density peaks are in the wrong places on the landscape. It is a 2BB library based purely on reductive aminations. The badly predicted DELs with the lowest Tc, were analyzed, but no trend was found in terms of the number of BBs, reactions, or degree of inclusion of their BBs in the train set. More examples of predicted maps for internal set and external set DELs are given in **Figure 7** and **Figure 8**.

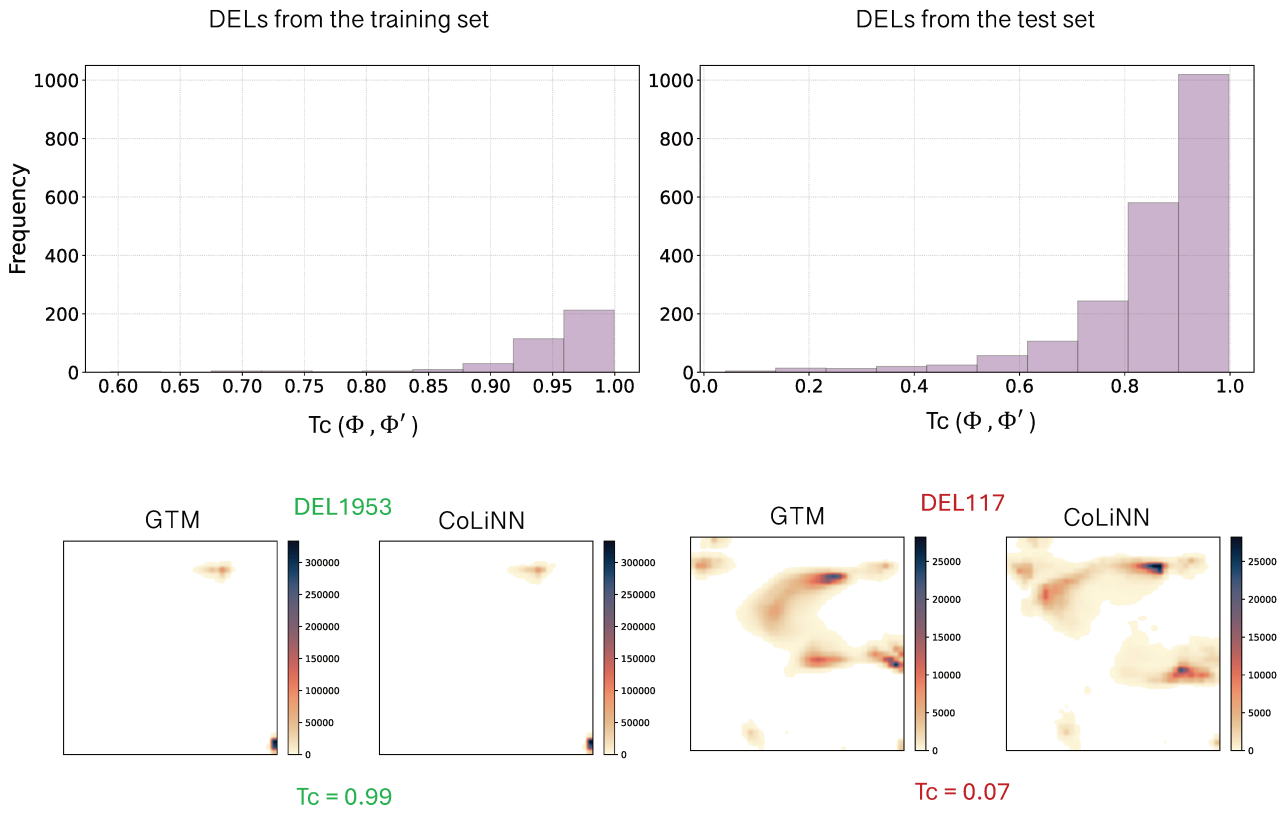


Figure 6. Histograms showing the distribution of Tc(ϕ, ϕ') for either training or test set DELs. Examples of badly and perfectly predicted DEL maps (from the test set) visualized as density landscapes are given below.

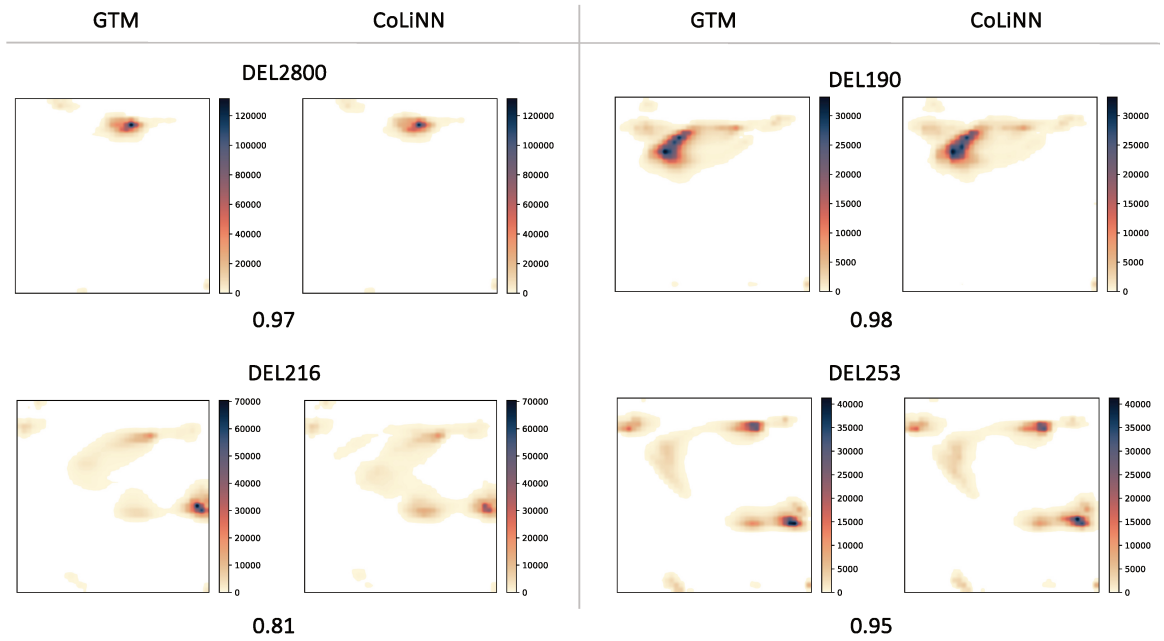


Figure 7. Density landscapes of four DELs that participated in the training, herein only 990K compounds absent from the training set were projected. Landscapes on the left are GTM-produced, the ones on the right were predicted by CoLiNN. DEL216 and DEL190 are 2BB DELs, all others are 3BB libraries. The values underneath the two maps correspond to the Tc similarity between them.

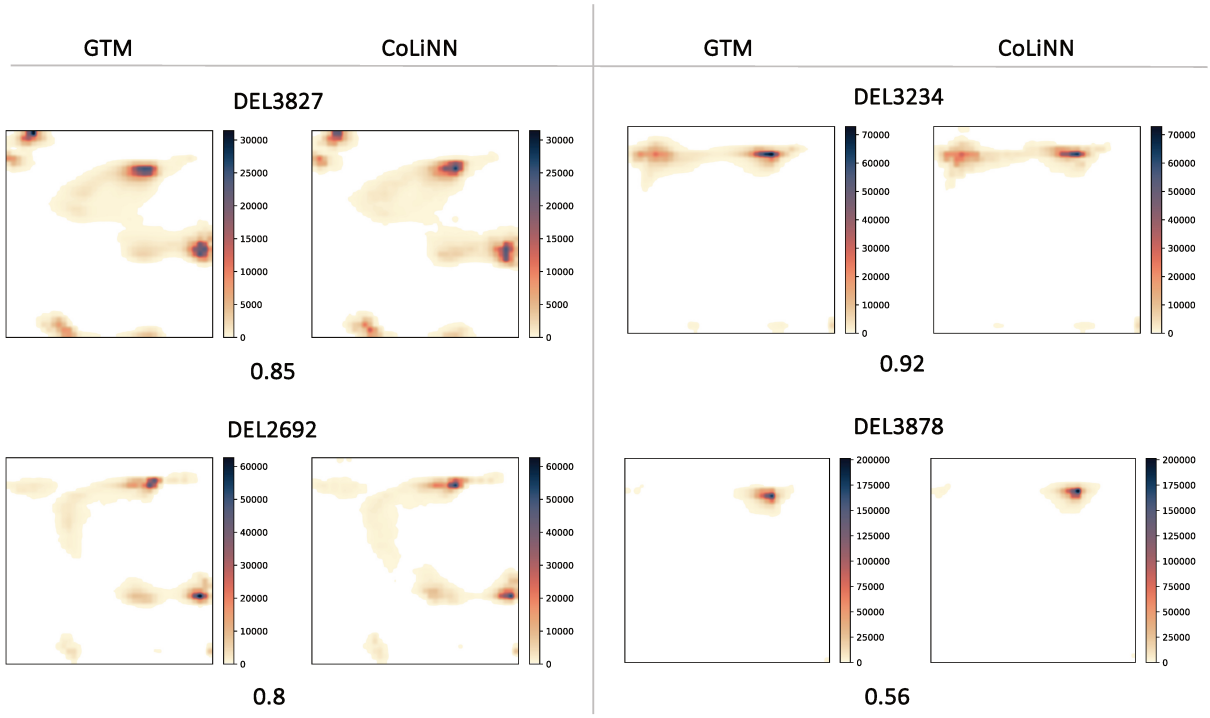


Figure 8. Density landscapes of four external test set DELs that did not participate in CoLiNN training, herein all 1M compounds per DEL are projected. Landscapes on the left are GTM-produced, the ones on the right were predicted by CoLiNN. All are 3BB libraries. The values underneath the two maps correspond to the Tc similarity between them.

To see the performance of the global CoLiNN model applied to the visualization of an ultra-large compound library, it was tested for prediction of the map of 80M-sized DEL2568. This particular DEL was not a part of 388 DELs used for global CoLiNN training and validation. The predicted density landscape of this DEL2568, shown in **Figure 9** mimics very well the true density landscape produced by the GTM algorithm. Tc between them is 0.91, showcasing the high accuracy of the global CoLiNN model for the prediction of even multimillion-sized combinatorial library maps.

Notwithstanding, the main idea behind CoLiNN is to predict correctly the "library chemical space motif" to be able to accelerate chemical library comparison. Hence, further on we will look at how well the ranking by similarity of DELs to ChEMBL (reference database) is preserved when using GTM-calculated or CoLiNN-predicted maps. For this, we calculated the Tc(ϕ_{DEL}, ϕ_{ChEMBL}) based on true GTM responsibilities and Tc(ϕ'_{DEL}, ϕ_{ChEMBL}) based on responsibilities predicted by CoLiNN, their distribution is given in **Figure 10** (a). To compare the rankings of DELs according to Tc similarity to ChEMBL obtained either by GTM or CoLiNN, Spearman ρ and Kendall τ correlation coefficients were used. The two rankings correlate as reflected by high Spearman ρ = 0.956 and Kendall τ = 0.828. The correlation between Tc values is also visually apparent from the joint distribution plot in **Figure 10** (a). However, what is more important here to look at is the closest DELs to ChEMBL with Tc=0.2-0.44. A zoomed joint distribution plot in **Figure 10** (b) shows that there are DELs for which the predicted similarity diverges from the true similarity. As an example, two such DELs (DEL532 and DEL1847) highlighted in red on the plot were analyzed. Their Tc values and density landscapes calculated using GTM or predicted by CoLiNN are given in **Figure 10** (d). For comparison purposes, the density landscape for ChEMBL28 is given in **Figure 10** (c). Their density landscapes clearly show that CoLiNN did not mispredict the library chemical space

motif – the GTM-calculated map and predicted one are very similar to each other. These results together clearly show that density landscapes predicted by CoLiNN provide almost the same ranking by similarity to a reference database (here ChEMBL) as true GTM-calculated ones. Such correspondence between rankings confirmed that there is no longer the need to enumerate combinatorial libraries to be able to analyse and compare them with respect to a reference database – CoLiNN allows to do it without.

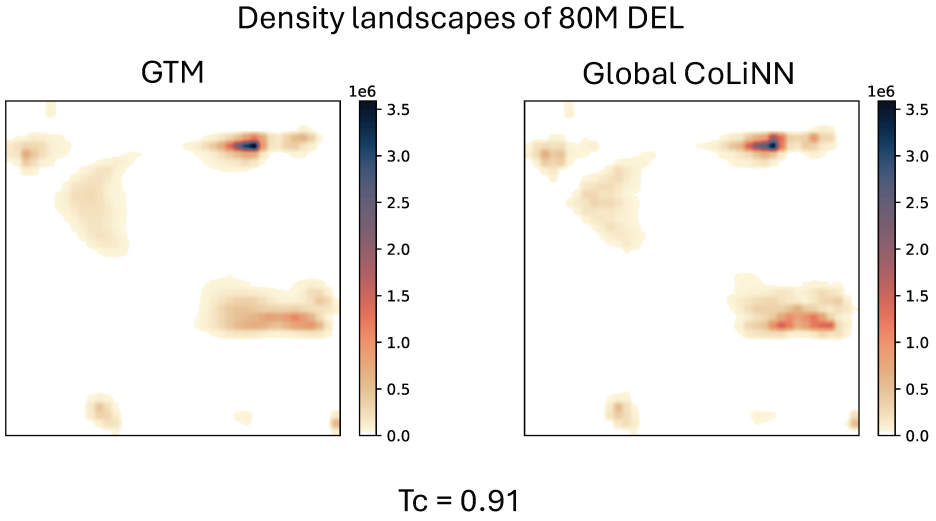
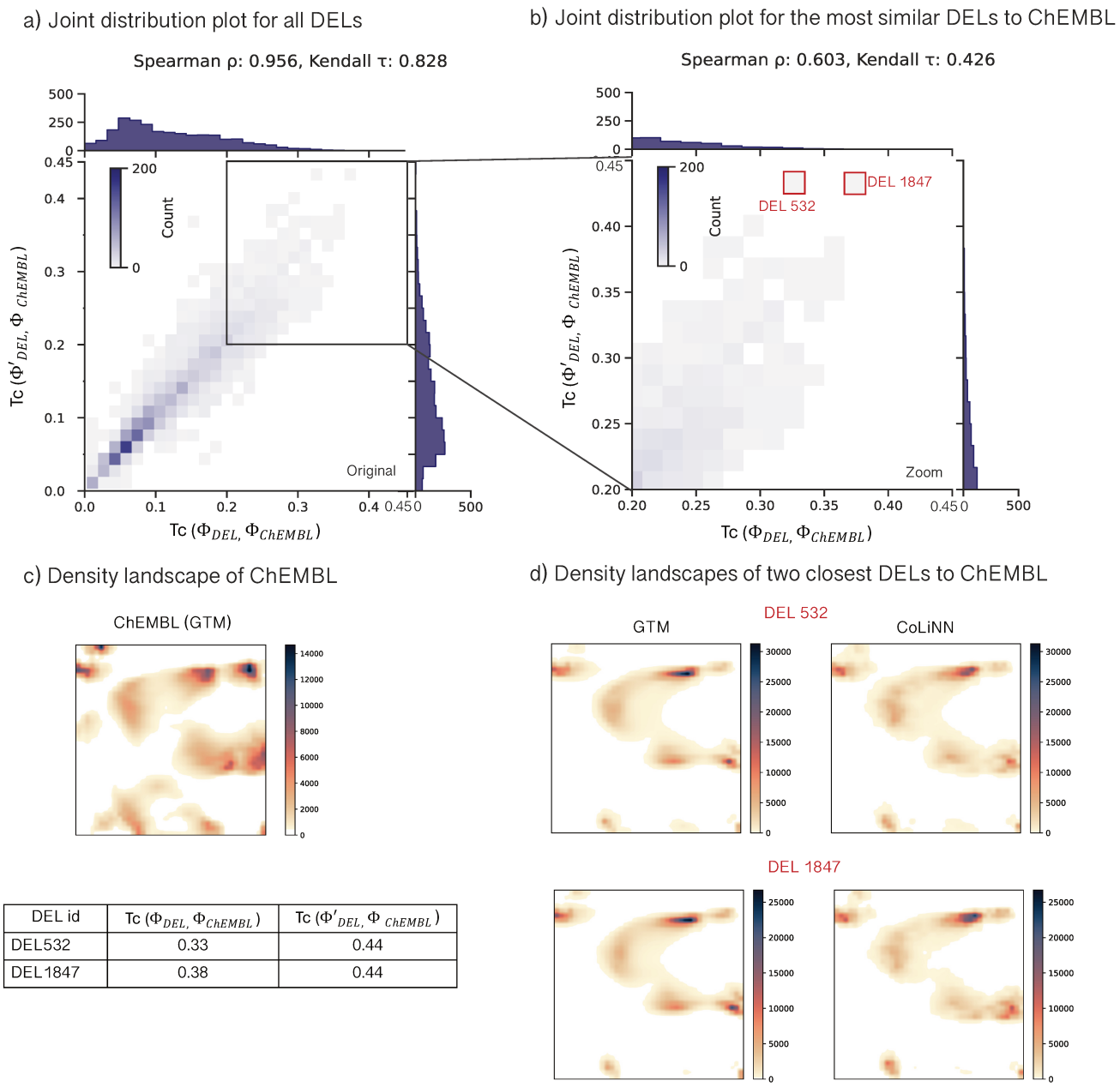


Figure 9. Density landscapes of ultra-large DEL2568 containing 80M compounds. On the left: density landscape calculated using GTM algorithm. On the right: density landscape predicted by Global CoLiNN.

CoLiNN vs GTM: calculation time

CoLiNN was trained on a single GPU of type NVIDIA RTX A6000, CUDA version 12.2, GPU memory 48 GB. The physical time needed to train global and local CoLiNN models as well as the physical prediction time expressed in ms/molecule/GPU are given in **Table 2** below. It should be noted that the major difference in time between global and local models is due to the fact that the mixed precision option was used for both training and inference of the global model. In more detail, mixed precision training and inference involve using both 16-bit and 32-bit floating-point types for computation, which can lead to increased speed and reduced memory usage.

The enumeration, standardization, descriptor calculation and GTM algorithm launch were performed on a machine powered by dual Intel Xeon Silver 4214R CPU at 2.40 GHz, total number of CPUs being 48. The typical time needed to perform all these calculations expressed in ms/molecule/CPU is given in **Table 3**. All CoLiNN models provide acceleration from 3000 to 7000-fold compared to traditional enumeration-based workflow. This shows that CoLiNN represents a better option in terms of calculation time/precision ratio for ultra-large combinatorial chemical space visualization when fragment-based descriptors and GTM dimensionality reduction methods are used.



DEL id	Tc ($\Phi_{DEL,}$ Φ_{ChEMBL})	Tc ($\Phi'_{DEL,}$ Φ_{ChEMBL})
DEL532	0.33	0.44
DEL1847	0.38	0.44

Figure 10. a) Joint distribution plot showing the correlation between Tc(ϕ_{DEL}, ϕ_{ChEMBL}) and Tc(ϕ'_{DEL}, ϕ_{ChEMBL}), as well as individual Tc distributions on the margins. The brightness represents the density of data points (counts). b) A zoomed version of the joint distribution plot showing the correlation between Tc(ϕ_{DEL}, ϕ_{ChEMBL}) and Tc(ϕ'_{DEL}, ϕ_{ChEMBL}) for the closest DELs to ChEMBL, two DELs that show less correlation are highlighted in red. d) Density landscape of ChEMBL28 and a table showing the Tc values for the two highlighted DELs. d) Density landscapes of the two highlighted DELs, on the left: GTM landscapes, on the right: landscapes predicted by CoLiNN.

Table 2. Physical time spent for training and validation of different CoLiNN models (including calculation of graphs and processing of building block and reaction indices) as well as total physical prediction time using these models.

Model	Total physical training time	Total physical prediction time (ms/molecule/GPU)
Global (388 DELs):		
3M 880K cmpds	13 h	0.055
Local:1M cmpds	14 h 37 m	0.124
Local: 50K cmpds	47 m 35 s	0.091
Local: 25K cmpds	23 m 17 sec	0.103
Local: 10K cmpds	21 m 8 sec	0.103

Table 3. Physical time spent on each step of the traditional enumeration-based workflow of compound visualization on the GTM.

Task	Physical time (ms/molecule/CPU)
Enumeration using eDesigner*	0.24
Standardization	187.2
Descriptor calculation	207.36
Projection	0.288
Total	395.08

*Runs only on a single process

CONCLUSION

In this work, we present the CoLiNN graph convolutional network trained to predict GTM-produced chemical space maps of ultra-large combinatorial libraries without compound enumeration. Two CoLiNN models (local and global) were trained. A global model was trained on 388 DELs employing 65K building blocks and 29 reactions and within 50 epochs showed high predictive performance. A local model was trained one DEL2568. As a performance metric, we used the Tanimoto similarity coefficient (Tc) calculated between the predicted and true cumulated GTM projection vectors. The global model was tested on the 2089 DELs that did not participate in training but shared BBs and reactions with the 388 training set libraries. For 1600 out of 2089 DELs, the predicted projection patterns have high similarity to the ones generated by the GTM algorithm – the Tc(ϕ, ϕ') spanned 0.8-0.99. Eventually, since cumulated GTM projection vectors are used to compare the chemical space coverage of libraries, predicted projection vectors are acceptable if they can effectively replace the "true" ones and yield similar results in ranking DELs by their similarity to the reference database, for example. Therefore, we compared the rankings of DELs based on their degree of similarity with ChEMBL chemical space using both "true" and predicted GTM maps. The rankings of DELs according to their similarity to ChEMBL remained highly consistent when replacing true DEL projection vectors with those predicted using CoLiNN (Spearman's $\rho = 0.956$). True, general CoLiNN predictions were deceiving with a few rare DELs, and it is still unclear why those failures happened, as these were neither the most complex nor the most "original" DELs (in the sense of poor BB and reaction scheme overlap with training data). However, for the practical design of a DEL based on a chosen chemistry, training a dedicated CoLiNN approach would avoid such glitches, and selection of the optimal BB set allowing for a tailor-made chemical space coverage should be feasible with important gains in speed.

These results showcase the high performance of CoLiNN in predicting the combinatorial chemical space motif projected on the GTM without compound enumeration. This tool can be used by medicinal chemists for library design accelerating the process of testing different building block combinations, visualization of ultra-large combinatorial chemical spaces, etc. As an improvement to the current CoLiNN

architecture in the future, the reactions may be represented by Condensed Graphs of Reaction[27], extending the application of CoLiNN to new unseen reactions. At the same time, the CoLiNN can be applied not only to predict GTM projection but also for other dimensionality reduction methods. In addition, prospectively, CoLiNN can be implemented not only for visualization but for the prediction of any property of combinatorial compounds, for example, docking score.

DATA AND SOFTWARE AVAILABILITY

The Python code of CoLiNN is available in the following GitHub repository:

https://github.com/Laboratoire-de-Chemoinformatique/CoLiNN

The data used in this work are available in the public domain resources: biologically relevant compounds from ChEMBL[28] (version 28) – https://www.ebi.ac.uk/chembl/, collection of eMolecules[13] building blocks partially available on the site https://www.emolecules.com/products/building-blocks, collection of Enamine[14] building blocks available on the site https://enamine.net/building-blocks.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] R. Liu, X. Li, K. S. Lam, *Curr Opin Chem Biol* **2017**, *38*, 117–126.

[2] B. Suay-García, J. I. Bueso-Bordils, A. Falcó, G. M. Antón-Fos, P. A. Alemán-López, *Int J Mol Sci* **2022**, *23*, DOI 10.3390/ijms23031620.

[3] R. Pikalyova, Y. Zabolotna, D. M. Volochnyuk, D. Horvath, G. Marcou, A. Varnek, *Mol Inform* **2022**, *41*, 2100289.

[4] R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, *J Chem Inf Model* **2023**, *63*, 4042–4055.

[5] R. Pikalyova, Y. Zabolotna, D. Horvath, G. Marcou, A. Varnek, *J Chem Inf Model* **2023**, *63*, 5571–5582.

[6] R. A. Goodnow, *A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery*, Wiley, **2014**.

[7] M. Rarey, J. S. Dixon, *J Comput Aided Mol Des* **1998**, *12*, 471–490.

[8] J. Lübbers, U. Lessel, M. Rarey, *J Chem Inf Model* **2024**, *64*, 2008–2020.

[9] L. Bellmann, P. Penner, M. Rarey, *J Chem Inf Model* **2021**, *61*, 238–251.

[10] L. Bellmann, P. Penner, M. Gastreich, M. Rarey, *J Chem Inf Model* **2022**, *62*, 553–566.

[11] L. Bellmann, R. Klein, M. Rarey, *J Chem Inf Model* **2022**, *62*, 2800–2810.

[12] D. K. Agrafiotis, V. S. Lobanov, *J Comput Chem* **2001**, *22*, 1712–1722.

[13] "eMolecules Inc. https://www.emolecules.com/2020.," **n.d.**

[14] "Enamine Ltd. https://enamine.net/2020.," **n.d.**

[15] R. A. Lerner, S. Brenner, *Angewandte Chemie International Edition* **2017**, *56*, 1164–1165.

[16] A. Martín, C. A. Nicolaou, *Commun Chem* **n.d.**, 1–9.

[17] C. M. Bishop, M. Svensén, C. K. I. Williams, *Neural Comput* **1998**, *10*, 215–234.

[18] D. Horvath, G. Marcou, A. Varnek, *Drug Discov Today Technol* **2020**, *xxx*, 1–9.

[19] Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, A. Varnek, *J Chem Inf Model* **2021**, *61*, 179–188.

[20] T. Akhmetshin, A. Lin, D. Mazitov, Y. Zabolotna, E. Ziaikin, T. Madzhidov, A. Varnek, *J Chem Inf Model* **2022**, *62*, 3524–3534.

[21] "Indigo Toolkit," **n.d.**

[22] M. Fey, J. E. Lenssen, *arXiv preprint arXiv:1903.02428* **2019**.

[23] D. Hendrycks, K. Gimpel, **2016**.

[24] T. N. Kipf, M. Welling, **2016**.

[25] M. Lee, *Journal of Mathematics* **2023**, *2023*, 4229924.

[26] A. H. Huang, **2024**.

[27] F. Hoonakker, N. Lachiche, A. Varnek, A. Wagner, *Int. J. Artif. Intell. Tools* **2011**, *20*, 253–270.

[28] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res* **2012**, *40*, 1100–1107.

SUPPORTING INFORMATION

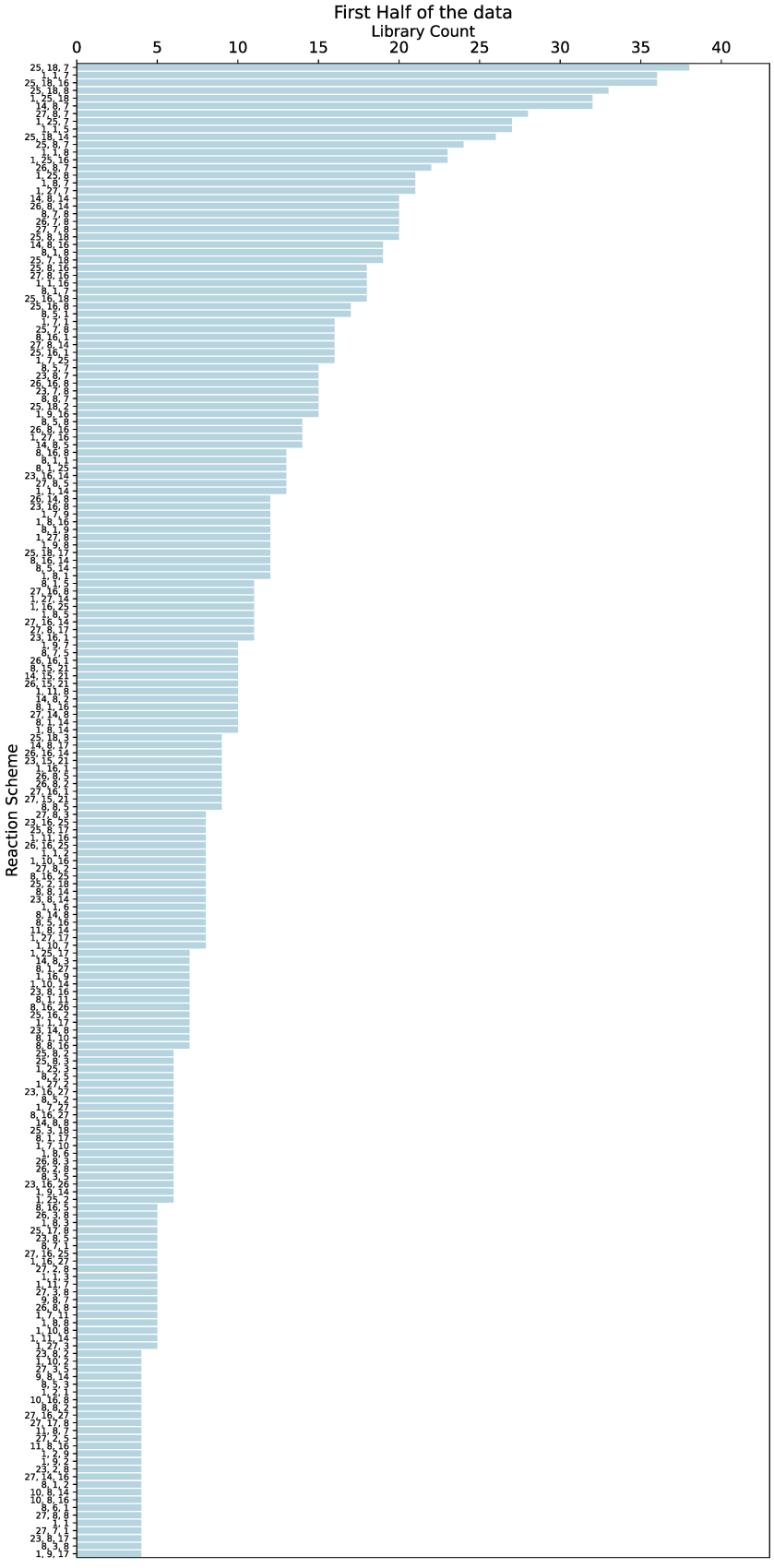


Figure S 1. Number of DELs with different reaction schemes. 1st Part of the data.

Figure S 2. Number of DELs with different reaction schemes. 2nd Part of the data.

Table S 1. Reactions used for enumeration of 2473 DELs used in this work. Their CoLiNN IDs, eDesigner numerical codes and reaction names as encoded in eDesigner are given.

ID CoLiNN	eDesigner numerical code	Reaction name encoded in eDesigner
1	1.2.1	Aldehyde_reductive_amination_FROM_aldehydes _AND_amines.rxn
2	1.3.1	Bromo_Buchwald-Hartwig_amination_FROM_arylbromide_AND_amines_a romatic.rxn
3	1.3.3	Iodo_Buchwald-Hartwig_amination_FROM_amines_aromatic_AND_aryli odide.rxn
4	1.3.8	Fluoro_N- arylation_FROM_amines_AND_nitro_fluoro.rxn
5	1.7.11	SNAr_ether_synthesis_FROM_phenols_AND_NAS_electrophil e.rxn
6	1.8.5	Thioether_synthesis_FROM_thiophenols_AND_NAS_electroph ile.rxn
7	1.8.7	Migita_thioether_synthesis_FROM_arylbromide_ AND_thiophenols.rxn
8	2.1.2	Carboxylic_acid_+_amine_condensation_FROM_amines_AND _carboxylic_acids.rxn
9	2.2.3	Sulfonamide_Schotten-Baumann_FROM_amines_AND_sulfonyl_chlorides.rxn
10	2.3.1	Isocyanate_+_amine_urea_coupling_FROM_amines_aliphatic_ AND_isocyanates.rxn
11	2.3.2	Isothiocyanate_+_amine_thiourea_coupling_FROM_amines_ali phatic_AND_isothiocyanates.rxn
12	3.1.1	Bromo_Suzuki_coupling_FROM_arylbromide_AND _boronics.rxn
13	3.1.3	Iodo_Suzuki_coupling_FROM_aryliodide_AND_bor onics.rxn
14	3.11.13	Ullmann-type_biaryl_coupling_FROM_amines_aliphatic_primary_AND_ arylbromide.rxn
15	3.9.21	Alkyne_coupling_FROM_alkynes_terminal_AND_ carboxylic_acids.rxn
16	4.1.1	1_2_3-Triazole_synthesis_FROM_azide_AND_alkynes_terminal. rxn
17	4.1.11	Larock_indole_synthesis_FROM_alkynes_terminal_AND_o_io do_aniline.rxn
18	4.1.12	Imidazole_synthesis_FROM_guanidines_tertiary_primary_AND _ketones_a_bromo.rxn
19	4.1.45	Benzimidazole_synthesis_FROM_aldehydes_AND _o_nitro_sec_aniline.rxn
20	4.1.48	Pyrimidine_synthesis_FROM_guanidines_sectert_primary_AN D_ynones.rxn
21	4.1.60	Pyrazole_synthesis_FROM_ynones_AND_hydrazi nes_primary.rxn

22	4.2.17	1,3-Benzoxazole_synthesis_FROM_amines_aliphatic_primary_AND_o_amino_phenols.rxn
23	4.2.2	1,2,4-Oxadiazole_synthesis_FROM_amines_aliphatic_primary_AND_carboxylic_acids.rxn
24	4.3.9	Benzothiazole_synthesis_FROM_amines_aliphatic_primary_AND_o_amino_thiophenols.rxn
25	9.7.8	Amino_to_guanidino_FROM_amines_aliphatic_AND_amines_aliphatic.rxn
26	X.X.X	Aminothiazole_synthesis_FROM_amines_aliphatic_primary_AND_ketones_a_bromo.rxn
27	X.X.X	SNAr_aniline_synthesis_FROM_amines_aliphatic_primary_AND_NAS_electrophile.rxn SNAr_aniline_synthesis_FROM_NAS_electrophile_AND_amines_aliphatic_secondary.rxn
28	X.X.X	SnAr_aniline_synthesis_FROM_triazine_chloro_AND_amines.rxn
29	X.X.X	Carboxylic_acid_+_nitro_condensation_FROM_carboxylic_acids_AND_o_nitro_sec_aniline.rxn

Table S 2. Values of the loss (KL divergence) at the beginning, halfway, and the end of training of local CoLiNN on a single DEL.

	Train loss	**Validation loss**
	1M subset	
Epoch 1	9.24	4.2
Epoch 50	0.34	0.35
Epoch 100	0.25	0.3
	50K subset	
Epoch 1	57.27	24.96
Epoch 50	2.57	2.66
Epoch 100	1.58	1.68
	25K subset	
Epoch 1	86.81	40.2
Epoch 50	3.82	3.76
Epoch 100	2.35	2.4
	10K subset	
Epoch 1	113.19	98.14
Epoch 50	4.52	4.52
Epoch 100	4.19	4.53

Summary

In this work, a GCN-based Combinatorial Library Neural Network (CoLiNN) model was developed to predict compound projections on the GTM using only their BBs and reactions. CoLiNN enables skipping the compound enumeration step, thereby accelerating the library analysis using GTM several thousand-fold compared to the enumeration-based workflow.

The local CoLiNN model, trained on small subsets of the 80M DEL allowed to accurately predict the map of the full library. It was determined that 25K compounds is an optimal training set size that is enough for CoLiNN to reach highly accurate predictions. The accuracy of these predictions was measured using the Tanimoto coefficient (Tc) between the Φ vectors of the predicted and the 'true' GTMs of the 80M DEL, yielding a Tc value of 0.96.

The general CoLiNN model, trained on 65K BBs and 29 reactions from 388 DELs, was also shown to accurately predict compound projection vectors on the GTM without the need for compound enumeration. It was tested on 2089 DELs that did not participate in training but shared BBs and reactions with 388 training set DELs. For 1600 DELs out of 2089, it allowed to predict the chemical space map with high accuracy. The latter was measured by the Tc between the Φ vectors of the predicted and “true” GTMs that spanned 0.8-0.99 for 1600 libraries. The global CoLiNN also showed high performance in predicting the chemical space map of the full 80M test set DEL. The Tc between the predicted and “true” Φ vectors for this library was 0.91. This result proves that the general CoLiNN allows to correctly predict the GTM of even ultra-large sized combinatorial libraries.

To evaluate whether the predicted by CoLiNN projection vectors can be used instead of the “true” ones generated by the GTM algorithm in the library comparison task, the ranking of DELs with respect to their similarity to the ChEMBL28 database was performed using both predicted and “true” Φ vectors. The two rankings showed a high correlation with Spearman $\rho = 0.956$. This result showcases that predicted by CoLiNN GTM projection vectors can indeed replace the “true” ones and still provide a consistent ranking of DELs according to their similarity to ChEMBL.

Overall, CoLiNN represents an efficient alternative method of compound library analysis that can predict chemical space maps of ultra-large DELs with high accuracy

without compound enumeration. Global CoLiNN predictions are 7000 faster than the typical workflow including enumeration, standardization, descriptor calculation, and GTM projection. Thus, CoLiNN holds potential for even non-enumerable combinatorial space visualization and analysis.

9. General conclusion and perspectives

DNA-Encoded Libraries (DELs) have emerged as an alternative way of hit identification allowing simultaneous screening of ultra-large chemical libraries inaccessible to classical high-throughput screening. The ultra-large size and the combinatorial nature of DELs bring unique challenges to their design, analysis, and comparison to other chemical libraries. This thesis addresses these challenges by developing efficient chemoinformatic methods of DEL chemical space exploration with a particular focus on interpretable machine learning approaches. At first, the limits of chemical space achievable through systematic enumeration of DELs using commercially available building blocks (BBs) were assessed by calculating the coverage of known biologically relevant compounds from the ChEMBL database by DELs. This was done in a low-dimensional space using Generative Topographic Mapping (GTM), which provides an illustrative way to compare chemical spaces. To speed up calculations, we developed an enumeration-free approach to chemical space visualization by combining GTM with deep neural networks, allowing scaling to ultra-large combinatorial chemical spaces. Additionally, we proposed a machine learning (ML) workflow for rational BB acquisition and hit prioritization tasks, based on the models predicting BB reactivity and DEL compound activity.

Large-scale generation and analysis of the DEL space

In this thesis, an ultra-large space of 2.5K DELs was designed using the eDesigner[11] tool from commercially available Building Blocks (BBs) and DNA-compatible reactions. Per DEL, a 1M representative subset of compounds was generated. Since Generative Topographic Mapping (GTM) has proven to be an efficient and accurate method of ultra-large chemical space visualization[9,15,47,52,69], it was used to visualize and analyze the space of DELs. A comparison of the GTMs of 2.5K DELs with ChEMBL28, a database of biologically relevant compounds, revealed that all DELs together cover the largest portion of chemotypes present in ChEMBL. However, DELs expectedly cannot cover regions of ChEMBL populated by natural products (NP), since in this work there was no intention to design NP-like DEL compounds. The GTM-based methodology allowed to identify pools of three and five complementary DELs that provided the highest coverage of biorelevant chemotypes from ChEMBL, 77% and 81%, respectively.

This project represents the first chemoinformatic generation and analysis of DELs at such a large scale, paving the way for further thorough investigation of DEL chemical space from various drug-discovery relevant perspectives. For instance, in the future, libraries generated in this study can be compared to the ZINC virtual screening collection of purchasable compounds to better understand the scope and value of the DEL space. With the increased availability of DEL-compatible BBs from sources like Mcule[98], Enamine[99], and Chemspace[100], the BB collection used here can be expanded to generate even more DELs and analyze them. Additionally, here, it was shown that many DELs cover similar chemical space regions. Therefore, future research could focus on enhancing the DEL design process by selecting BBs that result in structurally distinct libraries, minimizing redundant overlaps.

Efficient methods for the comparison of thousands of compound libraries

In this project, a Chemical Library Space (CLS) concept was introduced, which is composed of chemical libraries just as a Compound Space (CS) is composed of molecules. A vectorial representation of compound libraries based on the GTM was developed that defines the position of a compound collection within the CLS. This representation encodes the chemical space of an entire compound collection in a single vector, allowing it to be treated as an individual object and thus quickly compare chemical libraries. Using this methodology, 2.5K DELs were compared to ChEMBL either by structural or property similarity. It was revealed that the DELs that are the most similar to ChEMBL and thus more structurally diverse are all based on robust coupling reactions, whereas the majority of the most dissimilar DELs are based on heterocyclizations only.

The developed methodology was further extended by using a meta-GTM approach to produce a 2D map, where each chemical library is represented as a separate mapped object. This approach represents a valuable CLS visualization tool that allows one to get a “bird’s eye view” of the space composed of thousands of compound libraries. It was successfully applied to visualize the DEL CLS from different perspectives placing the libraries on the meta-map according to either their structural or property resemblance. The coloring of the meta-map by reaction types allowed to conclude that almost all the generated space could be covered by coupling-based DELs.

Comparing pairs of compound libraries using library vectors or visualizing all CLS on meta-GTM can significantly accelerate the selection of the optimal collection

based on chemotype, property coverage, or intrinsic library characteristics such as library size, synthesis cost, reactions used in synthesis, etc. This methodology is invaluable for narrowing down the search for an optimal compound library from thousands of possibilities, especially for specific drug discovery tasks that require the simultaneous consideration of multiple parameters. In the future, the developed chemical library space comparison approaches can be applied not only to DELs but to any other compound libraries. For example, tasks such as library diversification, selection of the analog collection with better synthesizability and cheaper reagents, and focused library design - can all be effectively supported by library vector comparison and analysis on meta-GTM. In the case of DELs, it would be interesting to see how the selection of a different BB set for its enumeration can shift its position in the CLS using meta-GTM.

BB reactivity prediction and hit prioritization: BRD4 focused DEL study

The GTM-based methodology was also shown to be effective for the analysis of the focused DEL which was experimentally tested against a BRD4 protein by Novalix. The coverage of the known space of BRD4 inhibitors from ChEMBL32 by hits from this focused DEL was estimated using GTM. Almost all hits were projected to the same areas as known BRD4 ligands with half-maximal inhibitory concentrations in the micromolar range confirming their affinity towards the protein. To further prioritize several hits for off-DNA synthesis a performant SVR model allowing to predict the pIC50(BRD4) values was trained on publicly available data on BRD4 inhibitors from ChEMBL32[95]. This model, coupled with GTM results, allowed to prioritize DEL hits from Novalix for further off-DNA synthesis. In addition, given the reactivity labels of the BBs used for the BRD4 focused DEL synthesis, highly performant SVM models for reactivity prediction were developed. These models will allow the future acquisition of only those BBs that were predicted reactive thus allowing to decrease the number of time-consuming and expensive BB reactivity validation experiments.

The SVM models developed herein in combination with GTM thus allowed to facilitate and interpret the results from the two major steps included in DEL technology - affinity screening and BB reactivity validation. These models can serve medicinal chemists for faster and more rational BB acquisition as well as for hit prioritization. One of the perspectives in focused DEL analysis could be the development of an end-to-end chemoinformatic platform accounting both for redundant and unreactive BBs exclusion,

chemical space coverage estimation, and ADME-Tox property prediction that could help evaluate the promise of DEL compounds in drug discovery campaigns.

CoLiNN as an efficient tool for combinatorial library visualization without compound enumeration

In the final project, the issue of compound enumeration, which limits the generation of combinatorial library compounds due to time and computational resource constraints, was addressed. A Combinatorial Library Neural Network (CoLiNN) was developed to predict compound projections on the GTM without the need for their enumeration. CoLiNN requires only the building blocks (BBs) and reactions used to create a compound as input. Trained on 388 DELs, the general CoLiNN model accurately predicted GTM projections for the majority of 2 089 DELs not included in the training set and achieved high accuracy in predicting the GTM projection of an ultra-large 80M compound DEL. This approach accelerates library visualization 7 000-fold compared to the conventional workflow, which includes compound enumeration, standardization, descriptor calculation, and GTM projection.

This project paves the way for the ultra-large combinatorial library visualization without the need for compound enumeration. The developed CoLiNN model is the first of its kind that is thoroughly tested on a substantial number of libraries utilizing different BBs and reactions. It was proven effective and accurate in predicting DEL compound projections on the GTM. Additionally, CoLiNN can be adapted to predict positions on the chemical space maps generated by any other dimensionality reduction method. Future enhancements could include using Condensed Graph of Reaction[101] (CGR) encoding to make CoLiNN's representation of reactions more universal, enabling predictions for combinatorial compounds with new reactions not included in the training set. The significance of reaction information can also be evaluated by excluding it from training and observing its impact on CoLiNN's performance. Moreover, the model can be tested on new combinatorial libraries such as the Enamine REAL combinatorial space. Another potential direction is to adapt CoLiNN for predicting other properties of combinatorial compounds.

10. Abbreviations

BA	Balanced Accuracy
BB	Building Block
BET	Bromodomain and Extraterminal protein family
BRD4	Bromodomain containing protein 4
CLS	Chemical Library Space
CoLiNN	Combinatorial Library Neural Network
CRV	Cumulated Responsibility Vector
DEL	DNA-Encoded Library
DL	Deep Learning
EF	Enrichment Factor
GA	Genetic Algorithm
GCN	Graph Convolutional Network
GTM	Generative Topographic Mapping
HTS	High Throughput Screening
IC50	Half-maximal inhibitory concentration
ML	Machine Learning
NCRV	Normalized CRV
PCR	Polymerase Chain Reaction
pIC50	Negative logarithm of the IC50 value

PKPD	Pharmacokinetic Pharmacodynamic
QED	Quantitative Estimate of Drug likeness
ROC AUC	Receiver Operating Characteristic Area Under the Curve
sEH	Soluble Epoxide Hydrolase
SVM	Support Vector Machine
SVR	Support Vector Regression
Tc	Tanimoto coefficient
UGTM	Universal Generative Topographic Mapping

11. References

(1) Brenner, S.; Lerner, R. A. Encoded Combinatorial Chemistry. *Proc Natl Acad Sci U S A* **1992**, *89* (12), 5381–5383.

(2) Lerner, R. A.; Brenner, S. DNA-Encoded Compound Libraries as Open Source: A Powerful Pathway to New Drugs. *Angewandte Chemie International Edition* **2017**, *56* (5), 1164–1165. https://doi.org/https://doi.org/10.1002/anie.201612143.

(3) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U. Impact of High-Throughput Screening in Biomedical Research. *Nat Rev Drug Discov* **2011**, No. March.

(4) Goodnow, R. A. *A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery*; Wiley, 2014.

(5) Ottl, J.; Leder, L.; Schaefer, J. V.; Dumelin, C. E. Encoded Library Technologies as Integrated Lead Finding Platforms for Drug Discovery. *Molecules* **2019**, *24* (8), 1–22. https://doi.org/10.3390/molecules24081629.

(6) Eidam, O.; Satz, A. L. Analysis of the Productivity of DNA Encoded Libraries. *Medchemcomm* **2016**, *7* (7), 1323–1331. https://doi.org/10.1039/C6MD00221H.

(7) Faver, J. C.; Riehle, K.; Lancia, D. R. Jr.; Milbank, J. B. J.; Kollmann, C. S.; Simmons, N.; Yu, Z.; Matzuk, M. M. Quantitative Comparison of Enrichment from DNA-Encoded Chemical Library Selections. *ACS Comb Sci* **2019**, *21* (2), 75–82. https://doi.org/10.1021/acscombsci.8b00116.

(8) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput* **1998**, *10* (1), 215–234. https://doi.org/10.1162/089976698300017953.

(9) Zabolotna, Y.; Lin, A.; Horvath, D.; Marcou, G.; Volochnyuk, D. M.; Varnek, A. Chemography: Searching for Hidden Treasures. *J Chem Inf Model* **2021**, *61* (1), 179–188. https://doi.org/10.1021/acs.jcim.0c00936.

(10) Pikalyova, R.; Zabolotna, Y.; Volochnyuk, D. M.; Horvath, D.; Marcou, G.; Varnek, A. Exploration of the Chemical Space of DNA-encoded Libraries. *Mol Inform* **2022**, *41* (6), 2100289.

(11) Martín, A.; Nicolaou, C. A. Navigating the DNA Encoded Libraries Chemical Space. *Commun Chem* No. 2020, 1–9.

(12) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, 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* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(13) Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Chemical Library Space: Definition and DNA-Encoded Library Comparison Study Case. *J Chem Inf Model* **2023**, *63* (13), 4042–4055.

(14) Pikalyova, R.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Varnek, A. Meta-GTM: Visualization and Analysis of the Chemical Library Space. *J Chem Inf Model* **2023**, *63* (17), 5571–5582. https://doi.org/10.1021/acs.jcim.3c00719.

(15) 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* (1), 84–94. https://doi.org/10.1021/ci500575y.

(16) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol Inform* **2010**, *29* (12), 855–868. https://doi.org/https://doi.org/10.1002/minf.201000099.

(17) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *J Med Chem* **2005**, *48* (7), 2518–2525.

(18) Jones, K. E.; Patel, N. G.; Levy, M. A.; Storeygard, A.; Balk, D.; Gittleman, J. L.; Daszak, P. Global Trends in Emerging Infectious Diseases. *Nature* **2008**, *451* (7181), 990–993. https://doi.org/10.1038/nature06536.

(19) Vasan, N.; Baselga, J.; Hyman, D. M. A View on Drug Resistance in Cancer. *Nature* **2019**, *575* (7782), 299–309. https://doi.org/10.1038/s41586-019-1730-1.

(20) Löscher, W.; Potschka, H.; Sisodiya, S. M.; Vezzani, A. Drug Resistance in Epilepsy: Clinical Impact, Potential Mechanisms, and New Innovative Treatment Options. *Pharmacol Rev* **2020**, *72* (3), 606–638. https://doi.org/10.1124/pr.120.019539.

(21) Moreau, J.; Mas, E. Drug Resistance in Inflammatory Bowel Diseases. *Curr Opin Pharmacol* **2015**, *25*, 56–61. https://doi.org/https://doi.org/10.1016/j.coph.2015.11.003.

(22) Li, A. P. Screening for Human ADME/Tox Drug Properties in Drug Discovery. *Drug Discov Today* **2001**, *6* (7), 357–366. https://doi.org/https://doi.org/10.1016/S1359-6446(01)01712-3.

(23) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J Comput Aided Mol Des* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(24) Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-Encoded Chemistry: Enabling the Deeper Sampling of Chemical Space. *Nat Rev Drug Discov* **2017**, *16* (2), 131–147. https://doi.org/10.1038/nrd.2016.213.

(25) Wellaway, C. R.; Amans, D.; Bamborough, P.; Barnett, H.; Bit, R. A.; Brown, J. A.; Carlson, N. R.; Chung, C.; Cooper, A. W. J.; Craggs, P. D.; Davis, R. P.; Dean, T. W.; Evans, J. P.; Gordon, L.; Harada, I. L.; Hirst, D. J.; Humphreys, P. G.; Jones, K. L.; Lewis, A. J.; Lindon, M. J.; Lugo, D.; Mahmood, M.; McCleary, S.; Medeiros, P.; Mitchell, D. J.; O'Sullivan, M.; Le Gall, A.; Patel, V. K.; Patten, C.; Poole, D. L.; Shah, R. R.; Smith, J. E.; Stafford, K. A. J.; Thomas, P. J.; Vimal, M.; Wall, I. D.; Watson, R. J.; Wellaway, N.; Yao, G.; Prinjha, R. K. Discovery of a Bromodomain and Extraterminal Inhibitor with a Low Predicted Human Dose through Synergistic Use of Encoded Library Technology and Fragment Screening. *J Med Chem* **2020**, *63* (2), 714–746. https://doi.org/10.1021/acs.jmedchem.9b01670.

(26) *X-CHEM: DNA-Encoded Library (DEL) Track Record of Success*. https://www.x-chemrx.com/del-track-record/ (accessed 2024-07-03).

(27) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; Cuozzo, J. W.; Guié, M.-A.; Guilinger, J. P.; Huguet, C.; Hupp, C. D.; Keefe, A. D.; Mulhern, C. J.; Zhang, Y.; Riley, P. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J Med Chem* **2020**, *63* (16), 8857–8866. https://doi.org/10.1021/acs.jmedchem.0c00452.

(28) *WuXi AppTec DELight*. https://discoverybiology.wuxiapptec.com/delight (accessed 2024-07-03).

(29) *Amgen: DNA-Encoded Libraries Will Drive Drug Design*. https://www.amgen.com/stories/2019/11/dna-encoded-libraries-will-drive-new-drug-design-paradigm (accessed 2024-07-03).

(30) Kontijevskis, A. Mapping of Drug-like Chemical Universe with Reduced Complexity Molecular Frameworks. *J Chem Inf Model* **2017**.

(31) *dynabind*. https://dynabind.com (accessed 2024-07-03).

(32) Petersen, L. K.; Christensen, A. B.; Andersen, J.; Folkesson, C. G.; Kristensen, O.; Andersen, C.; Alzu, A.; Sløk, F. A.; Blakskjær, P.; Madsen, D.; Azevedo, C.; Micco, I.; Hansen, N. J. V. Screening of DNA-Encoded Small Molecule Libraries inside a Living Cell. *J Am Chem Soc* **2021**, *143* (7), 2751–2756. https://doi.org/10.1021/jacs.0c09213.

(33) Blakskjaer, P.; Heitner, T.; Hansen, N. J. V. Fidelity by Design: Yoctoreactor and Binder Trap Enrichment for Small-Molecule DNA-Encoded Libraries and Drug Discovery. *Curr Opin Chem Biol* **2015**, *26*, 62–71. https://doi.org/https://doi.org/10.1016/j.cbpa.2015.02.003.

(34) Hansen, M. H.; Blakskjær, P.; Petersen, L. K.; Hansen, T. H.; Højfeldt, J. W.; Gothelf, K. V.; Hansen, N. J. V. A Yoctoliter-Scale DNA Reactor for Small-Molecule Evolution. *J Am Chem Soc* **2009**, *131* (3), 1322–1327. https://doi.org/10.1021/ja808558a.

(35) Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol Transl Sci* **2021**, *4* (4), 1265–1279. https://doi.org/10.1021/acsptsci.1c00118.

(36) Satz, A. L. What Do You Get from DNA-Encoded Libraries? *ACS Med Chem Lett* **2018**, *9* (5), 408–410. https://doi.org/10.1021/acsmedchemlett.8b00128.

(37) Halford, B. How DNA-Encoded Libraries Are Revolutionizing Drug Discovery. *Chem. Eng. News* **2017**, *95* (25), 28–33.

(38) Arico-Muendel, C. C. From Haystack to Needle: Finding Value with DNA Encoded Library Technology at GSK. *Medchemcomm* **2016**, *7* (10), 1898–1909. https://doi.org/10.1039/C6MD00341A.

(39) Lim, K. S.; Reidenbach, A. G.; Hua, B. K.; Mason, J. W.; Gerry, C. J.; Clemons, P. A.; Coley, C. W. Machine Learning on DNA-Encoded Library Count Data Using an Uncertainty-Aware Probabilistic Loss Function. **2021**.

(40) Eidam, O.; Satz, A. L. Analysis of the Productivity of DNA Encoded Libraries. *Medchemcomm* **2016**, *7* (7), 1323–1331. https://doi.org/10.1039/C6MD00221H.

(41) Franzini, R. M.; Neri, D.; Scheuermann, J. DNA-Encoded Chemical Libraries: Advancing beyond Conventional Small-Molecule Libraries. *Acc Chem Res* **2014**.

(42) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. V. S. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. *J Chem Inf Comput Sci* **2002**, *42* (2), 375–385. https://doi.org/10.1021/ci010375j.

(43) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing Novel Building Blocks Is an Overlooked Strategy to Improve Compound Quality. *Drug Discov Today* **2015**, *20* (1), 11–17. https://doi.org/https://doi.org/10.1016/j.drudis.2014.09.023.

(44) Fitzgerald, P. R.; Dixit, A.; Zhang, C.; Mobley, D. L.; Paegel, B. M. Building Block-Centric Approach to DNA-Encoded Library Design. *J Chem Inf Model* **2024**. https://doi.org/10.1021/acs.jcim.4c00232.

(45) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opin Drug Discov* **2015**, *10* (9), 959–973. https://doi.org/10.1517/17460441.2015.1060216.

(46) Lin, A.; Beck, B.; Horvath, D.; Marcou, G.; Varnek, A. Diversifying Chemical Libraries with Generative Topographic Mapping. *J Comput Aided Mol Des* **2019**, No. July. https://doi.org/10.1007/s10822-019-00215-x.

(47) Casciuc, I.; Zabolotna, Y.; Horvath, D.; Marcou, G.; Bajorath, J.; Varnek, A. Virtual Screening with Generative Topographic Maps: How Many Maps Are Required? *J Chem Inf Model* **2019**, *59* (1), 564–572. https://doi.org/10.1021/acs.jcim.8b00650.

(48) Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-Task Generative Topographic Mapping in Virtual Screening. *J Comput Aided Mol Des* **2019**, *33* (3), 331–343. https://doi.org/10.1007/s10822-019-00188-x.

(49) 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* (12), 1087–1108. https://doi.org/10.1007/s10822-015-9882-z.

(50) 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* (1), 84–94. https://doi.org/10.1021/ci500575y.

(51) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol Inform* **2015**, *34* (6–7), 348–356. https://doi.org/10.1002/minf.201400153.

(52) Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: A New Look at the Natural Product Chemical Space. *Mol Inform* **2021**, *40* (9), 2100068. https://doi.org/https://doi.org/10.1002/minf.202100068.

(53) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol Inform* **2015**, *34* (6–7), 348–356. https://doi.org/10.1002/minf.201400153.

(54) Orlov, A. A.; Khvatov, E. V.; Koruchekov, A. A.; Nikitina, A. A.; Zolotareva, A. D.; Eletskaya, A. A.; Kozlovskaya, L. I.; Palyulin, V. A.; Horvath, D.; Osolodkin, D. I.; Varnek, A. Getting to Know the Neighbours with GTM: The Case of Antiviral Compounds. *Mol Inform* **2019**, *38* (5), 1–12. https://doi.org/10.1002/minf.201800166.

(55) Mannocci, L.; Zhang, Y.; Scheuermann, J.; Leimbacher, M.; De Bellis, G.; Rizzi, E.; Dumelin, C.; Melkko, S.; Neri, D. High-Throughput Sequencing Allows the Identification of Binding Molecules Isolated from DNA-Encoded Chemical Libraries. *Proceedings of the National Academy of Sciences* **2008**, *105* (46), 17670–17675.

(56) Satz, A. L.; Hochstrasser, R.; Petersen, A. C. Analysis of Current DNA Encoded Library Screening Data Indicates Higher False Negative Rates for Numerically Larger Libraries. *ACS Comb Sci* **2017**, *19* (4), 234–238. https://doi.org/10.1021/acscombsci.7b00023.

(57) Xia, B.; Franklin, G. J.; Lu, X.; Bedard, K. L.; Grady, L. C.; Summerfield, J. D.; Shi, E. X.; King, B. W.; Lind, K. E.; Chiu, C.; Watts, E.; Bodmer, V.; Bai, X.; Marcaurelle, L. A. DNA-Encoded Library Hit Confirmation: Bridging the Gap Between On-DNA and Off-DNA Chemistry. *ACS Med Chem Lett* **2021**, *12* (7), 1166–1172. https://doi.org/10.1021/acsmedchemlett.1c00156.

(58) Favalli, N.; Bassi, G.; Scheuermann, J.; Neri, D. DNA-Encoded Chemical Libraries – Achievements and Remaining Challenges. *FEBS Letters*. Wiley Blackwell June 1, 2018, pp 2168–2180. https://doi.org/10.1002/1873-3468.13068.

(59) Fourches, D.; Tropsha, A. Using Graph Indices for the Analysis and Comparison of Chemical Datasets. *Mol Inform* **2013**, *32* (9–10), 827–842. https://doi.org/https://doi.org/10.1002/minf.201300076.

(60) 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 Cheminform* **2021**, *13* (1), 32. https://doi.org/10.1186/s13321-021-00505-3.

(61) Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M. Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. *J Chem Inf Model* **2022**, *62* (3), 553–566. https://doi.org/10.1021/acs.jcim.1c01378.

(62) Bellmann, L.; Penner, P.; Rarey, M. Connected Subgraph Fingerprints: Representing Molecules Using Exhaustive Subgraph Enumeration. *J Chem Inf Model* **2019**, *59* (11), 4625–4635. https://doi.org/10.1021/acs.jcim.9b00571.

(63) Shivanyuk, A. N.; Ryabukhin, S. V; Tolmachev, A.; Bogolyubsky, A. V; Mykytenko, D. M.; Chupryna, A. A.; Heilman, W.; Kostyuk, A. N. Enamine Real Database: Making Chemical Diversity Real. *Chemistry today* **2007**, *25* (6), 58–59.

(64) *CHEMriya Space*. Accessed 04.06.2024. https://www.otavachemicals.com/products/chemriya (accessed 2024-06-04).

(65) Detering, C.; Claussen, H.; Gastreich, M.; Lemmen, C. KnowledgeSpace - a Publicly Available Virtual Chemistry Space. *J Cheminform* **2010**, *2* (1), O9. https://doi.org/10.1186/1758-2946-2-S1-O9.

(66) *GalaXi space*. https://www.labnetwork.com/frontend-app/p/#!/ (accessed 2024-07-10).

(67) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol Inform* **2012**, *31* (3–4), 301–312. https://doi.org/10.1002/minf.201100163.

(68) Lin, A.; Horvath, D.; Marcou, G.; Beck, B.; Varnek, A. Multi-Task Generative Topographic Mapping in Virtual Screening. *J Comput Aided Mol Des* **2019**, *33* (3), 331–343. https://doi.org/10.1007/s10822-019-00188-x.

(69) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J. L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13* (6), 540–554. https://doi.org/10.1002/cmdc.201700561.

(70) Agrafiotis, D. K.; Lobanov, V. S. Multidimensional Scaling of Combinatorial Libraries without Explicit Enumeration. *J Comput Chem* **2001**, *22* (14), 1712–1722. https://doi.org/10.1002/jcc.1126.

(71) 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* (8), 1438–1454. https://doi.org/10.1021/acs.jcim.6b00192.

(72) Rodríguez-Pérez, R.; Vogt, M.; Bajorath, J. Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction. *ACS Omega* **2017**, *2* (10), 6371–6379. https://doi.org/10.1021/acsomega.7b01079.

(73) Rodríguez-Pérez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J Comput Aided Mol Des* **2022**, *36* (5), 355–362. https://doi.org/10.1007/s10822-022-00442-9.

(74) Whiteside, G. T.; Kennedy, J. D. Consideration of Pharmacokinetic Pharmacodynamic Relationships in the Discovery of New Pain Drugs. **2011**.

(75) Donati, B.; Lorenzini, E.; Ciarrocchi, A. BRD4 and Cancer: Going beyond Transcriptional Regulation. *Mol Cancer* **2018**, *17* (1), 164. https://doi.org/10.1186/s12943-018-0915-9.

(76) Hu, Y.; Zhou, J.; Ye, F.; Xiong, H.; Peng, L.; Zheng, Z.; Xu, F.; Cui, M.; Wei, C.; Wang, X.; Wang, Z.; Zhu, H.; Lee, P.; Zhou, M.; Jiang, B.; Zhang, D. Y. BRD4 Inhibitor Inhibits Colorectal Cancer Growth and Metastasis. *Int J Mol Sci* **2015**, *16* (1), 1928–1948. https://doi.org/10.3390/ijms16011928.

(77) Qian, H.; Zhu, M.; Tan, X.; Zhang, Y.; Liu, X.; Yang, L. Super-Enhancers and the Super-Enhancer Reader BRD4: Tumorigenic Factors and Therapeutic Targets. *Cell Death Discov* **2023**, *9* (1), 470. https://doi.org/10.1038/s41420-023-01775-6.

(78) Fish, P. V; Filippakopoulos, P.; Bish, G.; Brennan, P. E.; Bunnage, M. E.; Cook, A. S.; Federov, O.; Gerstenberger, B. S.; Jones, H.; Knapp, S. Identification of a Chemical Probe for Bromo and Extra C-Terminal Bromodomain Inhibition through Optimization of a Fragment-Derived Hit. *J Med Chem* **2012**, *55* (22), 9831–9837.

(79) Wyce, A.; Ganji, G.; Smitheman, K. N.; Chung, C.; Korenchuk, S.; Bai, Y.; Barbash, O.; Le, B.; Craggs, P. D.; McCabe, M. T. BET Inhibition Silences Expression of MYCN and BCL2 and Induces Cytotoxicity in Neuroblastoma Tumor Models. *PLoS One* **2013**, *8* (8), e72967.

(80) Gosmini, R.; Nguyen, V. L.; Toum, J.; Simon, C.; Brusq, J.-M. G.; Krysa, G.; Mirguet, O.; Riou-Eymard, A. M.; Boursier, E. V; Trottet, L. The Discovery of I-BET726 (GSK1324726A), a Potent Tetrahydroquinoline ApoA1 up-Regulator and Selective BET Bromodomain Inhibitor. *J Med Chem* **2014**, *57* (19), 8111–8131.

(81) Shapiro, G. I.; Dowlati, A.; LoRusso, P. M.; Eder, J. P.; Anderson, A.; Do, K. T.; Kagey, M. H.; Sirard, C.; Bradner, J. E.; Landau, S. B. Abstract A49: Clinically Efficacy of the BET Bromodomain Inhibitor TEN-010 in an Open-Label Substudy with Patients with Documented NUT-Midline Carcinoma (NMC). *Mol Cancer Ther* **2015**, *14* (12_Supplement_2), A49–A49.

(82) Siu, K. T.; Ramachandran, J.; Yee, A. J.; Eda, H.; Santo, L.; Panaroni, C.; Mertz, J. A.; Sims Iii, R. J.; Cooper, M. R.; Raje, N. Preclinical Activity of CPI-0610, a Novel Small-Molecule Bromodomain and Extra-Terminal Protein Inhibitor in the Therapy of Multiple Myeloma. *Leukemia* **2017**, *31* (8), 1760–1769.

(83) Albrecht, B. K.; Gehling, V. S.; Hewitt, M. C.; Vaswani, R. G.; Côté, A.; Leblanc, Y.; Nasveschuk, C. G.; Bellon, S.; Bergeron, L.; Campbell, R. Identification of a

Benzoisoxazoloazepine Inhibitor (CPI-0610) of the Bromodomain and Extra-Terminal (BET) Family as a Candidate for Human Clinical Trials. ACS Publications 2016.

(84) Hewitt, M. C.; Leblanc, Y.; Gehling, V. S.; Vaswani, R. G.; Côté, A.; Nasveschuk, C. G.; Taylor, A. M.; Harmange, J.-C.; Audia, J. E.; Pardo, E. Development of Methyl Isoxazoleazepines as Inhibitors of BET. *Bioorg Med Chem Lett* **2015**, *25* (9), 1842–1848.

(85) Taylor, A. M.; Vaswani, R. G.; Gehling, V. S.; Hewitt, M. C.; Leblanc, Y.; Audia, J. E.; Bellon, S.; Cummings, R. T.; Côté, A.; Harmange, J.-C. Discovery of Benzotriazolo [4, 3-d][1, 4] Diazepines as Orally Active Inhibitors of BET Bromodomains. *ACS Med Chem Lett* **2016**, *7* (2), 145–150.

(86) Gehling, V. S.; Hewitt, M. C.; Vaswani, R. G.; Leblanc, Y.; Côté, A.; Nasveschuk, C. G.; Taylor, A. M.; Harmange, J.-C.; Audia, J. E.; Pardo, E. Discovery, Design, and Optimization of Isoxazole Azepine BET Inhibitors. *ACS Med Chem Lett* **2013**, *4* (9), 835–840.

(87) Herait, P. E.; Berthon, C.; Thieblemont, C.; Raffoux, E.; Magarotto, V.; Stathis, A.; Thomas, X.; Leleu, X.; Gomez-Roca, C.; Odore, E. Abstract CT231: BET-Bromodomain Inhibitor OTX015 Shows Clinically Meaningful Activity at Nontoxic Doses: Interim Results of an Ongoing Phase I Trial in Hematologic Malignancies. *Cancer Res* **2014**, *74* (19_Supplement), CT231–CT231.

(88) Stathis, A.; Zucca, E.; Bekradda, M.; Gomez-Roca, C.; Delord, J.-P.; de La Motte Rouge, T.; Uro-Coste, E.; de Braud, F.; Pelosi, G.; French, C. A. Clinical Response of Carcinomas Harboring the BRD4–NUT Oncoprotein to the Targeted Bromodomain Inhibitor OTX015/MK-8628. *Cancer Discov* **2016**, *6* (5), 492–500.

(89) Gaudio, E.; Tarantelli, C.; Ponzoni, M.; Odore, E.; Rezai, K.; Bernasconi, E.; Cascione, L.; Rinaldi, A.; Stathis, A.; Riveiro, E. Bromodomain Inhibitor OTX015 (MK-8628) Combined with Targeted Agents Shows Strong in Vivo Antitumor Activity in Lymphoma. *Oncotarget* **2016**, *7* (36), 58142.

(90) McDaniel, K. F.; Wang, L.; Soltwedel, T.; Fidanze, S. D.; Hasvold, L. A.; Liu, D.; Mantei, R. A.; Pratt, J. K.; Sheppard, G. S.; Bui, M. H. Discovery of N-(4-(2, 4-Difluorophenoxy)-3-(6-Methyl-7-Oxo-6, 7-Dihydro-1 H-Pyrrolo [2, 3-c] Pyridin-4-Yl) Phenyl) Ethanesulfonamide (ABBV-075/Mivebresib), a Potent and

Orally Available Bromodomain and Extraterminal Domain (BET) Family Bromodomain Inhibitor. *J Med Chem* **2017**, *60* (20), 8369–8384.

(91) Seal, J.; Lamotte, Y.; Donche, F.; Bouillot, A.; Mirguet, O.; Gellibert, F.; Nicodeme, E.; Krysa, G.; Kirilovsky, J.; Beinke, S. Identification of a Novel Series of BET Family Bromodomain Inhibitors: Binding Mode and Profile of I-BET151 (GSK1210151A). *Bioorg Med Chem Lett* **2012**, *22* (8), 2968–2972.

(92) Mirguet, O.; Gosmini, R.; Toum, J.; Clément, C. A.; Barnathan, M.; Brusq, J.-M.; Mordaunt, J. E.; Grimes, R. M.; Crowe, M.; Pineau, O. Discovery of Epigenetic Regulator I-BET762: Lead Optimization to Afford a Clinical Candidate Inhibitor of the BET Bromodomains. *J Med Chem* **2013**, *56* (19), 7501–7515.

(93) Filippakopoulos, P.; Qi, J.; Picaud, S.; Shen, Y.; Smith, W. B.; Fedorov, O.; Morse, E. M.; Keates, T.; Hickman, T. T.; Felletar, I. Selective Inhibition of BET Bromodomains. *Nature* **2010**, *468* (7327), 1067–1073.

(94) Madsen, D.; Azevedo, C.; Micco, I.; Petersen, L. K.; Jakob, N.; Hansen, V. *An Overview of DNA-Encoded Libraries : A Versatile Tool for Drug Discovery*, 1st ed.; Elsevier B.V., 2020; Vol. 59. https://doi.org/10.1016/bs.pmch.2020.03.001.

(95) *ChEMBL32*. https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_32/ (accessed 2024-07-03).

(96) Pikalyova, K.; Orlov, A.; Lin, A.; Tarasova, O.; Marcou, M.; Horvath, D.; Poroikov, V.; Varnek, A. HIV-1 Drug Resistance Profiling Using Amino Acid Sequence Space Cartography. *Bioinformatics* **2022**, *38* (8), 2307–2314. https://doi.org/10.1093/bioinformatics/btac090.

(97) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. **2016**.

(98) *Mcule DEL-Compatible BB Library*. https://mcule.com/database/ (accessed 2024-07-03).

(99) *Enamine DEL-compatible scaffolds*. https://enamine.net/building-blocks/scaffolds (accessed 2024-07-03).

(100) *Chemspace Building Blocks for DEL*. https://chem-space.com/compounds/blocks-for-del (accessed 2024-07-03).

(101) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed Graph of Reaction: Considering a Chemical Reaction as One Single Pseudo Molecule. *Int. J. Artif. Intell. Tools* **2011**, *20* (2), 253–270.

Regina PIKALYOVA

Chémoinformatique des chimiothèques à codage ADN : design, génération in silico, gestion, analyse, et comparaison

Résumé

Cette thèse est dédiée à la génération de l'espace virtuel de 2.5K chimiothèques codées par ADN et à leur analyse chémoinformatique détaillée par structures et propriétés. Des méthodologies basées sur GTM permettant de comparer rapidement et de sélectionner les chimiothèques DEL optimales parmi des milliers de possibilités en fonction de la similarité structurelle ou par propriétés par rapport à une base de données de référence ont été développées. Le problème de l'énumération combinatoire des composés a été abordé en développant un modèle d'apprentissage profond capable de prédire la position d'un composé sur la GTM en se basant uniquement sur les réactifs associés et les réactions, en évitant l'énumération. Des modèles permettant de prédire la réactivité des réactifs et l'activité biologique des composés DEL ont été créés. Ces derniers couplés à la visualisation GTM, ont permis de rationaliser et de faciliter la sélection des réactifs pour la synthèse DEL et la priorisation des hits.

Mots-clés : Chimiothèques à codage ADN, chimiothèques combinatoires, comparaison de chimiothèques, GTM, visualisation de l'espace chimique, énumération des composés, apprentissage profond

Résumé en anglais

This thesis is dedicated to the generation of the virtual space of 2.5K DNA-encoded libraries and its detailed structural and property chemoinformatic analysis. GTM-based methodologies allowing to quickly compare and select optimal DELs out of thousands of possible ones based on the structural or property similarity to a reference database were developed. The problem of combinatorial compound enumeration was addressed by developing a highly performant deep learning model capable of predicting the position of a compound on the GTM based solely on its building blocks and reactions, skipping compound enumeration. Machine learning models allowing to predict the reactivity of building blocks and activity of DEL compounds were created. The latter coupled with GTM visualization, allowed to rationalize and facilitate the reagent selection for DEL synthesis and hit prioritization.

Keywords: DNA-encoded libraries, combinatorial libraries, library comparison, GTM, chemical space visualization, compound enumeration, deep learning