ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2024**, *97*(4)
Published online: November 30, 2024
DOI: 10.5562/cca4130

HKD 1926

The State of the Art on Chemical Databases and Libraries

Višnja Stepanić,[1,*] Dalibor Hršak,[1] Renata Kobetić[2]

[1] Division of Electronics, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

[2] Division of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

* Corresponding author's e-mail address: Visnja.Stepanic@irb.hr

RECEIVED: August 2, 2024 * REVISED: September 19, 2024 * ACCEPTED: September 19, 2024

THIS PAPER IS DEDICATED TO THE LATE PROFESSOR TOMISLAV CVITAŠ

Abstract: Molecules that act on the biological target at micromolar level at least are called hits. The usual method for identifying hits is high-throughput screening (HTS) of chemical libraries in relevant *in vitro* assays. An even more efficient, cost-effective and faster method for identifying hits is to perform virtual pre-screening, where the top scoring hits are validated in appropriate *in vitro* assays. Both wet HTS and virtual screening using structure- or ligand-based approaches utilise large libraries containing millions to billions of drug-like compounds. In this paper, we provide an insight into the state of the art in large collections of small molecular weight molecules, i) public databases for synthetic compounds (PubChem, ChEMBL) and natural products (COCONUT, LOTUS) and commercial ultra-large chemical libraries, ii) make-on-demand virtual libraries (Enamine, Galaxi®, ZINC-22) and iii) wet DNA-encoded libraries (DELs). Machine learning methods for characterising and visualising molecular diversity in screening collections are also described.

Keywords: database, library, ultra-large library, DNA-encoded library, virtual screening, hits, machine learning, visualization, PCA, t-SNE, UMAP.

INTRODUCTION

WE live in an era of extensive use and development of automated screening platforms including high-throughput screening (HTS) assays, aimed at the rapid, cost-effective and efficient discovery of hit molecules with targeted activity. Hits are low molecular weight (MW) molecules that show at least μM activity on a target macromolecule of interest (Figure 1).[1] Despite the increasing focus of large pharmaceutical and biotech companies on biological therapeutics like antibodies, vaccines and gene editing therapies, efforts to discover small MW compounds with specific biological activity continue in small and medium-sized companies.[2]

Projects in the life sciences are aimed to discover and develop modulators of the activities of biological target macromolecules, such as inhibitors, agonists or antagonists of disease-relevant proteins, or to develop fluorescent molecular probes to monitor certain biological processes. Although the chemical space is estimated to have 10^{63} organic molecules with up to 30 C, N, O and S atoms, only about 2650 small drugs (MW up to 850) are approved in the US, Canadian and EU markets.[3,4]

In the search for novel biologically active small molecules, various methods have been developed to optimise the efficiency of the discovery process, *i.e.* to reduce time and costs while mitigating the high risks associated with targeting complex biological networks. Nowadays, the design of novel molecules with a specific biological mechanism of action usually begins with the identification of hit molecules by HTS and their validation by appropriate *in vitro* counter-screens and orthogonal assays, whereby inactive compounds are filtered out and dismissed (Figure 1 a).[1] The first step of such HTS campaigns commonly involves multi-stage virtual screening (VS) of large chemical libraries (Figure 1 b). VS enables fast identification of a set of structurally diverse hit molecules with (sub)micromolar affinity among millions of compounds. *In silico* screening is based on the sequential application of different computational approaches including molecular modelling methods and unsupervised and supervised machine learning (ML) methods.

Unsupervised ML approaches include *e.g.* clustering of compounds and/or targets according to their structural similarity. Classification and quantitative structure activity relationship (QSAR) ML models are used for filtering out non-drug-like and inactive molecules.[6] Drug-likeness is a commonly applied concept with the aim to reduce a vast chemical space to those compounds that are similar to drugs in terms of absorption, distribution, metabolism and excretion (ADME) properties. It is commonly assessed by applying simple classification rules pioneered by Lipinski's rule-of-five (Ro5) for oral bioavailability.[5,6] Depending on the availability of the 3D structure of the target macromolecule and diverse active molecules, VS includes structure- and ligand-based methods such as molecular docking and pharmacophore modelling, respectively.[7] The ultimate goal of VS is to detect best-scoring and diverse hits which are further validated *in vitro* in relevant assays.

Large chemical libraries used for HTS are compiled either from proprietary collections or from commercially available compound collections by applying different *in silico* approaches and tools and taking into account the needs of a project. VS precedes *in vitro* screening as it can be efficiently applied to libraries containing billions of actual compounds within a few days on high-performance computing (HPC) clusters.[8] Structure-based VS not only reduces the number of compounds to be experimentally screened, but has also been shown to improve the hit rate in a screening by 10- to 1000-fold.[9] Pre-elimination of compounds with undesirable molecular features such as reactive electrophiles and redox active compounds or those containing toxicophores reduces the false positive rate of HTS and improves the optimization performance and cost of the downstream pipeline (Figure 1 a).

This review focuses on the state-of-the-art collections of compounds that have already been synthetized or can be readily made-on-demand. Such collections are available for researchers to design their own experiments/projects while searching for starting compounds and their analogues and/or target activities by VS. First, we give a brief overview of the publicly available chemical databases with data on millions of known chemicals. We then describe chemical libraries and their types. There are various sorts of chemical libraries that are designed according to their intended use. Thereafter, we present huge virtual libraries of synthesisable compounds that are made on demand and finally DNA-encoded libraries (DELs). DELs are collections of enumerated molecules, each of which is coupled with unique DNA tags that serve as amplifiable identification barcodes, allowing fast identification of active compounds. The review also describes ML methods for the description of chemical libraries, including their visualisation.

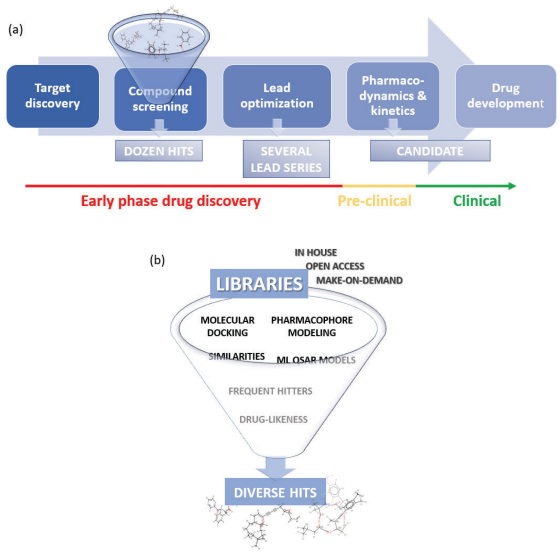

Figure 1. a) Modern drug discovery scheme. Screening of chemical libraries is carried out at a very early stage of drug discovery to identify hit compounds with an activity of at least 10 µM on a selected biological target.[9] Optimisation of the activity by modifying the structures of the hits generates lead compounds whose structures are further modified to optimize their pharmacodynamic and pharmacokinetic profile, ultimately leading to a candidate molecule that enters clinical trials. b) The VS cascade is composed of several computational steps with the aim of identifying the best-scored compounds – hits that are further validated *in vitro*.

PUBLICLY AVAILABLE CHEMICAL DATABASES

Chemical libraries can be ensembled from compounds which have already been synthetized and tested in biological assays, and are listed in an online database. The most commonly used freely available and manually curated databases are PubChem, ChEMBL and DrugBank (Table 1). These databases provide information on chemical, physicochemical, bioactivity and omics data as well as patents. They can be searched by compound structure, name or CAS number for similar compounds and biological activities, or by the a protein or gene name for the active and inactive compounds. Open-access chemical databases are frequently used for retrieving molecules with targeted biological activities using programmed scripts. The databases adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles, ensuring that the data they contain can be easily found, retrieved, integrated, and used for research purposes. Here we give a short introduction.

PubChem is a public chemical database launched by National Institutes of Health (NIH) in 2004.[10] It allows users to search for compounds, substances and BioAssay data. A compound is a normalized chemical structure representation found in one or more contributed substances. Given a particular chemical or protein/gene, PubChem is typically used by a synthetic chemist or molecular biologist to summarise a whole range of information about it by linking and integrating knowledge from various data sources including EPA DSSTox, DrugBank, Human Metabolome Database (HMDB), KEGG, Wikipedia, ChEMBL, *etc*. In other words, PubChem provides the integration of information on physicochemical and spectroscopic features, biological activities and targets and related patents as well as links to the databases from which they are retrieved. As of March 2024, PubChem contains data on 118 million (M) compounds, 315 M substances, 294 M bioactivities and 51 M patents, drawn from 983 data sources. Instructions for using PubChem online, downloading data via FTP protocol and programmatic access can be found at: https://pubchem.ncbi.nlm.nih.gov/docs/about.[11] PubChem also allows community participation by uploading their own data.

Launched in 2009 by the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), ChEMBL is a chemogenomic database that brings together chemical, bioactivity and genomic non-proprietary data to facilitate the translation of genomic information into new medicines.[12] All data are fully traceable and explicitly referenced to the original source. It contains 2.4 M drug-like compounds (accessed March 2024) and their binding, functional, ADME and toxicological bioactivity data (1.6 M bioassays; 5 K data with assigned mechanism of action for compounds) extracted from the primary scientific literature. In addition to small bioactive compounds, ChEMBL also contains peptides and therapeutic antibodies. Chemical Entities of Biological Interest (ChEBI) corresponds to ChEMBL's resources that focus on small chemicals, their nomenclature, structure, and biological properties.[13] ChEMBL - Neglected Tropical Disease (ChEMBL – NTD) is a subset of compounds targeting neglected tropical diseases in Africa, Asia and the Americas. SureChEMBL contains 14 M compounds extracted daily from 24 M patent documents including full texts, images and attachments of patent documents (accessed May 2024).[14] ChEMBL and ChEBI are part of the ELIXIR Core Data Resources. ELIXIR is a European intergovernmental organisation that coordinates and develops an infrastructure for the life sciences that provides biomolecular data, computational tools, training material, cloud storage and supercomputers (https://elixir-europe.org/services).[15]

Table 1. Links to frequently used online databases and make-on-demand chemical libraries.

Open-access databases	
PubChem	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL	https://www.ebi.ac.uk/chembl/
DrugBank	https://go.drugbank.com/
COCONUT	https://coconut.naturalproducts.net/
LOTUS	https://lotus.naturalproducts
TCM ID	https://bidd.group/TCMID/index.html
KNApSAcK	http://www.knapsackfamily.com/knapsack_core/top.php
Make-on-demand libraries	
ZINC20	https://zinc20.docking.org/
ZINC-22	https://cartblanche.docking.org/
Enamine REAL	https://enamine.net/
Chemspace Freedom	https://chem-space.com/compounds/freedom-space
WuXi AppTec GalaXi®	https://wuxibiology.com/drug-discovery-services/hit-finding-and-screening-services/
ULTIMATE	https://ultimate.mcule.com/

Since its release in 2006, DrugBank has grown to become a 'gold standard' resource for knowledge about drugs, their indications, physicochemical properties (including MS and chromatographic properties) and various pharmacodynamic and pharmacokinetic information including drug-drug and drug–target interactions.[4] The newest version DrugBank 6.0 contains data on 4563 FDA-approved drugs (a 72 % increase from the 2018 version), 6231 investigational drugs, 1 413 413 drug–drug interactions, 2475 drug–food interactions and 29802 drug–target interactions. An investigational drug is a medication in the development phase that is being studied for a specific condition and has entered clinical trials.

In addition to databases of molecules synthetized in labs, more than 120 electronic resources on natural products (NPs) have been published in the last 20 years.[16] NPs are often difficult to synthesize and cover different and more diverse areas of the chemical space than synthetic compounds.[17] NPs are secondary metabolites from living organisms, mostly microbes, plants and fungi. The development of new technologies greatly facilitates the development of NPs and their use especially as drugs, and increases the rate of unambiguous identification of new NPs from biological matrices from years to days. Three years ago, the first comprehensive open accessed online database COlleCtion of Open Natural prodUcTs (COCONUT) was published by Steinbeck's group.[18] COCONUT is the largest collection of 407 270 unique NPs mainly from 53 sources. Researchers from various fields, from drug discovery to research of molecular aspects of biodiversity can use the user-friendly COCONUT web interface (Table 1)

to search NPs in multiple ways: by their structure, compound name and simple molecular features related to drug-likeness and structural complexity. Structural complexity is usually described by the fraction of tetrahedral, sp^3 hybridised carbons Fsp3 and a number of stereogenic carbons. Increasing Fsp3 has been reported to improve solubility as well as target specificity and selectivity.[19] However, molecules with a high Fsp3 content are also generally more challenging to produce.

The LOTUS initiative is an upgrade of COCONUT with open assessed information on more than 750 000 NP-containing organism pairs and thus it represents a unique resource for taxonomic and evolutionary studies.[20] The LOTUS data is hosted in the community-managed knowledge base Wikidata which allows mining of the literature underlying the experimental work, and is also mirrored on the interactive web portal (Table 1) which allows searching for NPs in a similar way to COCONUT. COCONUT and LOTUS Initiative also contain compounds from traditional Chinese medicine (TCM). A representative TCM database is the TCM Information Database (TCM ID) which contains comprehensive information on all aspects of TCM including 7443 recipes, 2751 constituent herbs and 7375 herbal ingredients linked to therapeutic and side effects, putative targets and biological pathways.[21] The KNApSAcK is a comprehensive database of species-metabolite relationships that enables searches for metabolites and associated plant species and *vice versa* using integrated metabolite-plant species resources.[22]

CHEMICAL LIBRARIES

Chemical libraries are collections of compounds that are prepared for screening experiments carried out with the ultimate goal of discovering biologically active molecules, called hits (Figure 1). They can be made of proprietary, public or commercially available chemicals.[23] In general, screening libraries contain drug-like molecules, or certain chemical classes that lie outside the chemical space of Ro5 such as macrocycles or peptides. A library is assembled prior to synthesis from compounds/building blocks that have specific substructures or functionalities and associated specific molecular and/or biological features.

There are different sorts of libraries with respect to their generation strategies, domain of interest and commercial availability. Chemical libraries can be formed and enlarged by enumeration of side chains from central scaffolds such as privileged scaffolds which are known to generate biologically active compounds.[24,25] Examples of privileged structures yielding ligands for several biological receptors are benzodiazepines, indoles, quinolines, isoquinolines, benzoxazoles, coumarin, prostanoic acid, *etc*. Nowadays, combinatorial libraries can be generated by chemical assembly of a large set of building blocks, and such ultra-large libraries encompass a vast chemical space. The foundations of combinatorial chemistry were laid in 1963 by Merrifield in his report on the synthesis of peptides by addition of protected amino acids to a chain bound to a solid resin particle.[26] This approach enabled not only straightforward creation of desired sequences, but also simple purification of the peptide product by dissolving and washing away impurities instead of recrystallization. The mid to late 80s witnessed a strong development in combinatorial peptide chemistry with the advent of several new methods for combinatorial synthesis of peptides.[27] Although combinatorial libraries can be very large, the efforts to produce and test them often suffered from problems with assay artefacts due to limited solubility and a lack of diversity. A strategy using so-called DNA encoded libraries (DELs) has overcome some of these deficits.

Commercially available libraries are presented in form of various catalogues available on internet (*e.g.* https://enamine.net/compound-libraries). There are diversity, targeted, focused, fragment libraries as well as make-on-demand libraries. Make-on-demand compounds are synthetized on request of a customer and they usually correspond to hit molecules pre-selected by performing an extensive VS of (ultra-)large chemical libraries or their rationally chosen subsets (Figure 1 b). Make-on-demand libraries are formed virtually by applying computational methods for combinatorial generation of compounds taking into account various molecular features and synthesis possibilities. Such virtual libraries with compounds of high synthetic accessibility can be built from collection of starting building blocks with validated chemical reactions.[28]

Drug-like molecules are selected based on their similarity in terms of physicochemical and simple structural features with drugs and generally *in vivo* active compounds. These properties determine absorption *i.e.* oral bioavailability (blood concentration of a molecule after taking it *per os*), distribution, metabolism and excretion (ADME) of chemicals in the body, *i.e.* they largely determine their pharmacokinetic profile. Drug-like molecules fulfil statistically found rules in terms of simple structural features (*e.g.* numbers of hydrogen bond acceptor (HBA) and donor (HBD) atoms, topological polar surface area (TPSA)) and/or physicochemical features (lipophilicity, solubility and permeability).[5,29] Drug-likeness concept is successfully applied by medicinal chemists in the early stages of drug discovery to reduce the attrition rate in late clinical phases (Figure 1).[30] Biologically active macrocycles with MW greater than 500 have been shown to have a drug-like physicochemical profile similar to small compounds that fulfil Ro5.[31]

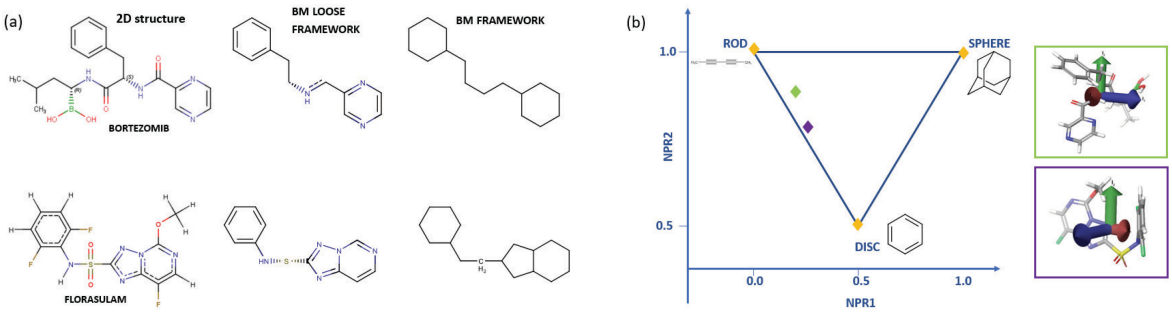

Figure 2. Characterisation of the chemical diversity of bortezomib and florasulam by (a) 2D BM scaffold variety and (b) 3D inertial ratios. Both chemicals have a rod-like shape as shown by their placement in the triangular rod/disk/sphere plot. NPR1 – 1st normalised PMI (*i.e.* I_1 / I_3), NPR2 – 2nd normalised PMI (*i.e.* I_2 / I_3).

For a large multi-million/billion collection of drug-like compounds, so-called **diversity library** corresponds to its reliable structural representation. Diversity library is a compilation of compounds that are diverse in their molecular structures or targets whose activity they modulate, and is intended for hit searching. **Targeted library** is designed as a screening collection of drug-like compounds directed at specific biological targets, including protein families such as protein kinases, G protein-coupled receptors, ion channels, proteases, epigenetic and protein-protein interaction related targets, nuclear receptors, and others. **Focused libraries** are compilations of molecules based on their structural similarity to compounds that have been shown to act against the target of interest. Higher hit rates are observed when screening target focused libraries than when screening diversity sets, and the resulting hit clusters typically exhibit SARs that facilitate subsequent structural optimisation.[32] Screening collections can also be made of compounds with potential biological activity against specific diseases (such as anticancer, antiviral or antibacterial collections), as well as specific areas of application (agrochemical libraries, targeted CNS library). As already mentioned, there are also collections of compounds with specific chemical classes such as peptides or macrocycles. Catalogues of different types of screening libraries can be found on the websites of companies offering diverse and pharmacologically relevant collections of drug-like or lead-like small molecules, such as Enamine's REAL, MCule's ULTIMATE, Chemspace's Freedom, WuXi's GalaXi®, TargetMol and so on. There are also **fragment libraries** of structurally diverse small MW chemicals which are built in a way to enable efficient exploration of chemical space and whose linking may produce a good lead molecule. The fragments usually satisfy the rule of 3 stating that on average, fragment hits tend to exhibit MW < 300, have ≤ 3 HBDs, ≤ 3 HBAs, a lipophilicity coefficient CLogP ≤ 3, number of rotatable bonds on average ≤ 3 and polar surface area around 60 Å^2.[33]

Structural Diversity

One way to characterize the structural novelty of chemical libraries is to assess their structural diversity. Diversity analysis helps to explore the heterogeneity of chemical libraries and to design diverse libraries for screening or SAR analysis.[34] Library diversity is commonly described in terms of scaffold diversity and shape diversity which are usually represented by a frequency count of 2D Bemis–Murcko (BM) scaffolds and the inertial ratios calculated from 3D conformations, respectively (Figure 2). The BM framework or skeleton represents a molecular backbone composing of ring systems and linkers connecting them (Figure 2a).[35] Compared to the BM framework/skeleton, the BM loose framework/scaffold also retains information on atom/bond types. The BM frameworks and scaffolds are usually used for clustering of compounds according to their 2D structural similarity and the number of clusters is used as a measure of diversity. Shape diversity is quantified by normalized ratios of first and second principal moments of inertia (PMI) which are calculated from 3D conformations and plotted into 2D triangular rod/disk/sphere graph (normalized principal moment of inertia ratio, NPR-analysis) (Figure 2b).[36,37] Molecular diversity can also be quantified using similarity metrics among which the Tanimoto coefficient (TC) is the most commonly used.[38] Molecules are usually represented with fingerprints which describe the presence (1) or absence (0) of atoms, bonds and various structural features and the TC is calculated by using fingerprint representation of compounds. The TC ranges from 0 for completely dissimilar molecules to 1 for identical molecules. Calculations of various diversity/similarity parameters are illustrated in Figure 2 on the example of two selected compounds - the drug bortezomib and the herbicide florasulam. 2D BM scaffolds are generated by Marvin (Figure 2a).[39] Like many structures and core scaffolds relevant to medicinal chemistry, these two compounds tend to have a rod-like shape (Figure 2b). The PMIs were calculated from rotational constants for 3D conformations of minima determined at the M062X/6-31+G** level by using Gaussian 16.[36,40] Measured by the TC with the representation of molecules by the MACCS fingerprints, bortezomib and florasulam are dissimilar compounds with the TC of only 0.361 (calculated with the R packages *rcdk* ver. 2.9).[41,42]

Visualization of Chemical Space

The chemical space is made up of compounds with different structures and physicochemical and biological features. Visualization provides a fast and efficient way to gain insight into the diversity of compounds in the library and the importance of structural features and/or physicochemical properties governing target activity. Various methods have been developed for visualization of chemical space. The most commonly used methods for visualising chemical space are Principal Component Analysis (PCA), t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).[43,44] All three methods reduce the dimensionality of data.

In PCA, the original set of n vectors is linearly transformed into a new set of k orthogonal principal components, while retaining as much information (variance) as possible from the original data set, *i.e.* the global structure of the data. In contrast, t-SNE and UMAP are non-linear dimensionality reduction methods. T-SNE is a randomised algorithm for embedding high-dimensional data in a two- or three- dimensional space while preserving the local structure of the data, *i.e.* the distance. T-SNE converts high-dimensional distances/similarities into conditional probabilities by minimising the Kullback–Leibler divergence between the distributions in the high- and low-dimensional space. For this purpose, the hyperparameters perplexity, the learning rate (eta) and the number of iterations are adjusted. Perplexity is related to the number of nearest neighbours and as a rule of thumb has a value of approximately $\sqrt{n}$, where n is the number of data points. Eta determines how fast the algorithm converges, and its starting value is usually 200. If the data is not well separated, one can try to increase eta, whereas if it is too scattered, it should be decreased. UMAP preserves both the local and global structure of the data by constructing a high-dimensional graph representing the manifold structure of the data, and then optimizing a low-dimensional graph to be as structurally similar as possible. UMAP parameters, a number of nearest neighbours and minimum distance can be tuned to balance between local and global structure preservation. Various distance metrics can be used to quantify similarities between data points in t-SNE and UMAP like Euclidean (usually the default metric), Manhattan, Jaccard (equals to 1-TC) and so on. Both t-SNE and UMAP can use PCA to initialize embedding.

The 2D visualizations obtained by PCA, t-SNE and UMAP for the chemical space of drugs are shown in Figure 3. The drugs were represented by the 166-bit MACCS fingerprints (more than 10 drugs with specific structural feature) calculated with the R package *rcdk*.[41,42] PCA and t-SNE were done by the R function *prcomp* and the package *Rtsne*, respectively, while UMAP was performed by DataWarrior ver. 06.02.01.[45,46] Jaccard distance was used for both methods t-SNE and UMAP. The best embedding of drugs according to their therapeutic use (ATC classes) is achieved by UMAP, as it is known to preserve local and global structure and relationships of data.[47]

Make-on-Demand Chemical Libraries

Recently, chemical screening libraries have grown to billions of molecules with some private commercial collections containing more than 10^{20} compounds, whereas in 2020 the number of compounds in stock worldwide was around 14 million.[23,48] Such ultra-large virtual compound collections which can be made on demand, greatly facilitate findings of diverse hits for novel targets as well as the discovery of new chemotypes for known targets as they enable rapid and cost-effective exploration of chemical space by VS.[7,49] They are ground-breaking as they not only increase the probability of finding hit molecules (at least due to their size), but also significantly reduce the time and

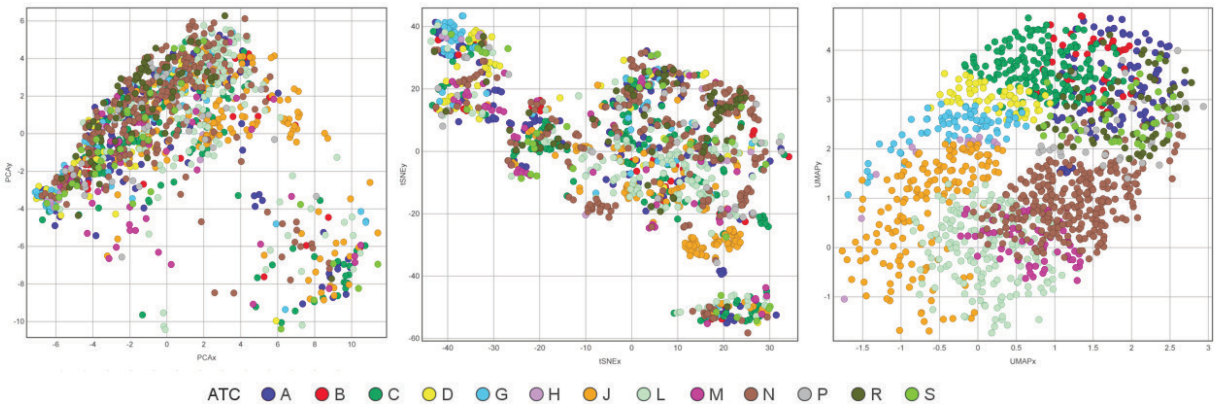

Figure 3. Comparison of 2D plots generated using PCA (left), t-SNE (middle) and UMAP (right) methods for the set of 1291 approved drugs represented by MACCS fingerprints and coloured according to their ATC classes A-S for therapeutic use. Plots are drawn in DataWarrior.

cost of synthesis and the time needed to optimise the structure and generate SARs for the biological target and other relevant biological parameters. The chemical diversity, scaffold diversity, and shape diversity of make-on-demand libraries far exceed public physical libraries, and probably private ones as 97 % of the scaffolds in make-on-demand libraries have no representative in physical libraries.[50] Here we list the most commonly used make-on-demand libraries (Table 1).

ZINC is a free enumerated database of commercially available and make-on-demand compounds established already in 2005.[51] Originally, it focused on providing 3D conformations of commercial *i.e.* synthesisable compounds in proper protonation and tautomeric forms for VS based on molecular docking. It is an analog-by-catalog (ABC), that enables fast search for similar compounds by using 2D atomic-level graph-based methods for whole-molecule similarity (SmallWord algorithm) and pattern and substructure search (Arthor), and exploring SARs. ZINC has grown from fewer than 1 M compounds in 2005 to 1.4 billion (B) compounds in 2020, of which 1.3 B can be acquired from 310 catalogues from 150 companies.[50] ZINC-22 provides an online search for more than 37 B enumerated commercially available compounds from the Enamine REAL Database (5 B), Enamine REAL Space (29 B), WuXi (2.5 B), Mcule (Ultimate 128M), and ZINC20 in stock (4 M).[52] Due to the huge increase in purchasable chemical space, compared to ZINC20, ZINC-22 has been restructured addressing challenges in disk access, rapid lookup, database distribution and download, and the relational database structure. There are more than 4.5 B 3D conformations ready-to-dock. Over 95 % of the available molecules with up to 24 heavy atoms (HAC24) and over 80 % with up to 25 heavy atoms (HAC25) have been built in 3D. The database is chemically and structurally diverse, reflecting the enormous effort by vendors to add new reaction schemes and, particularly, new building blocks.

Ukrainian company Enamine provides access to the largest enumerated virtual database of 48 B compounds that are REadily AccessibLe (REAL) through validated parallel synthesis using in-stock building blocks.[53] The REAL compounds are generated by more than 167 well-validated parallel synthesis protocols applied to over 143 000 qualified reagents and building blocks. The REAL space can be searched online for analogue compounds by using virtual chemical space navigation platform infiniSee developed by BioSolveIT (https://enaminestore.com/search)[54] Instead of TC, the infiniSee tool uses pharmacophoric features as a measure of similarity, which are not rigidly bound to substructure patterns but rather relate to molecular properties.

The further extension of the REAL space in terms of chemical scaffolds and their diversity is represented by the Freedom space.[55] The Freedom space with 5B make-on-demand compounds provided by Chemspace was developed according to the principles of Enamine's REAL space and consists of chemicals with a synthesis success rate of over 75 per cent, which differ from those of the REAL library and also the ChEMBL database.

WuXi AppTec GalaXi®is an online platform that provides a reaction-based ultra-large virtual library of 3.4 B compounds that complements the commercial chemical space. Launched in 2019, it is also the result of a strategic partnership with BioSolveIT, where WuXi AppTec provides selected and novel building blocks with proven chemistry and BioSolveIT contributes with its infiniSee software and algorithms.

The online platform mcule.com offers a huge space with around 6 M in-stock and purchasable chemicals which can be easily searched by structure and filtered by molecular features (https://mcule.com/database).[56] Mcule has also launched the virtual collection ULTIMATE with more than 100 M novel indexed chemicals generated by an AI algorithm ARCHIE to ensure an average synthetic feasibility rate of over 80 % by encoding robust, well-validated reactions rigorously reviewed by experts.[57] ULTIMATE can be searched by Pharmit based on pharmacophore, molecular shape, and energy minimization to enable scaffold hopping, navigating out of the patented space, and identifying new chemotypes.[58]

DNA-Encoded Libraries

DNA-Encoded Libraries (DELs) were first introduced in 1992 as encoded combinatorial chemical libraries.[59] In such libraries, the potential compounds of interest are tagged with unique DNA sequences that serve as their identifiers. When creating DELs, a building block is appended to a larger molecular structure, followed by an addition of an oligonucleotide sequence (the encoder) which serves as a tag for that particular building block. DELs harness the advantages of genetic methods, namely the possibility of amplification of DNA sequences using polymerase chain reaction (PCR), their enrichment with active molecules and easy decoding of active molecule structures by DNA sequencing, as well as the possibility of expanding the chemical space to millions of DNA tagged compounds using combinatorial methods.[60,61] In such a way, DELs facilitate identification of molecules that bind macromolecular targets.

At first, DELs were synthesised using the split-and-pool strategy.[62,63] The splitting part consisted of separating DNA oligonucleotides attached to an insoluble solid support material such as polystyrene beads into multiple vessels, while the pooling part included attaching a series of small molecules onto the oligonucleotide encoders. After each pooling the separated support beads

were mixed and split again in a random fashion. The differentiation between the DNA synthesis phase and the active molecule (peptide) synthesis phase was achieved by using orthogonal protection groups, that had to be attached to the reactive functional groups of either oligonucleotide or peptides.[64] Posterior selective deprotection enables the conjugation step without cross-reactivity.

The solution-phase synthesis offered a different approach by avoiding the constraints of solid-phase systems. This method allows for more straightforward purification processes and can be more suitable for synthesizing molecules that change their binding properties from solid phase to solution. The solution-phase synthesis also enabled the development of encoded self-assembling chemical (ESAC) libraries, which are synthesised through the aggregation of macromolecular units such as DNA into larger supramolecular complexes *via* non-covalent bonding (*e.g.* π-π stacking, hydrogen bonding, hydrophobic interactions) under appropriate conditions (temperature, pH, ion strength, *etc.*).[65] Such supramolecular structures spontaneously form both on solid surfaces and in solution. ESAC libraries can be large in size and diversity, as they can be constructed from smaller sub-libraries, and their oligonucleotide tags can form DNA-heteroduplexes and triplexes.

Large variety of binding ligands in an ESAC library has been achieved through the design of dual- and trio-pharmacophore DELs where two and three pre-purified sub-libraries, respectively, have been combined.[66,67] In trio-DEL, a fragment from one sub-library (denoted as SL-B) binds to the target and it serves as a linker to connect to fragments from the other two sub-libraries (SL-A and SL-C) through self-assembly of their flanking DNA strands.[67] The authors have designed a library with over 23 M unique members, developed a strategy for optimising the linkers using bovine trypsin and identified a series of *de novo* compounds with inhibitory activity on human matrix metalloproteinases. The applicability, size and robustness of dynamic dual display DELs has recently been further increased by a large encoding design (LED) strategy, with Y-shaped dynamic self-assembling DELs consisting of two partially complementary DNA strands, each bearing two sets of molecular building blocks.[68] The Y-junction splits the DNA strand into two sets of encoding regions, thus enabling separate and independent encoding sequences for each set of molecular assemblies and constructing DELs with a "2+2" configuration.

The early problem of combinatorial DEL chemistry for small molecules was the absence of straightforward strategies for amplifying the selected molecules of interest. The mere affinity selection, namely the enrichment of the library mixture through the iterative removal of the non-bound members resulted in relatively scarce hits among the vast number of molecules from the initial concoction. The first strategy for *in vitro* selection and evolution of combinatorial libraries was reported and patented in 2004.[69] The strategy consisted of three steps: 1) conversion of genes to their respective products, 2) product selection and 3) gene amplification, where the support material for the translation of genes to active molecules was a single-stranded DNA chain (ssDNA).

The first systematic study of discovery of new small MW ligands using large-scale DELs was reported in 2009, where double stranded DNAs (dsDNA) were used as encoders.[70] The dsDNA design provided enhanced tag stability during the synthetic processes and reduced likelihood of interference with the target. A library containing around 800 M different molecules produced using the split-and-pool method, was tested for inhibition against kinases p38 MAP and Aurora A, whereby three potential high affinity inhibitors were identified through affinity selection and enrichment measurement, whose SAR was in line with previously reported results. The method of affinity selection enables the identification of not only specific high affinity ligand molecules, but rather families of structurally related ligands.

DNA Encoded Dynamic Libraries (DEDLs) or Dynamic Combinatorial Libraries (DCL) originate from the concept of reversible thermodynamically controlled synthesis of macromolecular blocks.[71–73] In such libraries, the selection events stabilise the library members of interest and shift the chemical equilibrium towards their creation and away from the less desired members.[74] Such selection events enable a more efficient isolation of desired compounds with higher chemical yields. The first step in screening DEDLs/DCLs is the preparation of a mixture of molecules of interest, followed by the amplification of molecules that bind best to the target by shifting the chemical equilibrium, and the final step includes isolation of such binding molecules.[72] Fine tuning of conditions can be used for switching off the interconversion between potential binders. Non-covalently assembled binder-target complexes can further be stabilised through the reversible covalent binding.[75] This target-accelerated combinatorial synthesis has been successfully applied for the discovery of novel vancomycin derivatives as antibiotics against vancomycin-resistant bacterial strains, and of *in situ* assembly of acetylcholinesterase inhibitors (click-chemistry).[76,77] A similar approach named “extended tethering” for identifying small MW ligands for cysteine containing proteins has been reported by Erlanson *et al.*[78] Their method is based on reactions of a library of disulphide containing molecules with an unpaired cysteine-containing target protein under partially reducing conditions (with presence of 2-mercaptoethanol) that promote rapid thiol

exchange. By changing the redox state of the buffer, either thiols or disulphides can be favoured. If a library member shows inherent affinity for the protein, the equilibrium will shift toward the modified protein. A more recent view on dynamic DELs proposed principles of Darwinian evolution pressure to the selection of important compounds, by enabling the translation of DNA genetic sequences into synthetic molecules, instead of simply using them as mere tags.[79] The authors have addressed the issue of efficient screening of a vast chemical space that are beyond the reach of classical DELs, which extends to 10^{60} small drug-like molecules. By applying the Darwinian selection processes (affinity selection, panning), the DEL techniques, especially the dynamic ones, allow for finding hit molecules that are not present in the starting library.

One attractive alternative to tagging of active molecules using DNA was shown to be peptide (or polyamide) nucleic acid (PNA) tags, where the deoxyribose-phosphate backbone of DNA is replaced with a backbone based on thiamine-aminoethyl glycil.[80,81] Such PNA strands have a much stronger affinity towards their complementary DNA strand and such a PNA-DNA complex is more stable than a DNA-DNA double helix. A PNA-Based Dynamic Combinatorial Library (PDCL) has been used to screen the series of fucose-based glycans for their binding affinity to fucose binding sites in bacterial lectins.[82] The cooperative effects with PNA tags increased the affinity compared to using glycans alone.

Besides DELs being applied in finding ligands with strong binding affinities to proteins, the same principles can be applied for the discovery of ligands to other types of molecules, such as RNA.[83] Such targets can be desirable, because most proteins are not amenable for inhibition due to *e.g.* conformational reasons, so the researchers have pursued the route of targeting their encoding mRNA, thus inhibiting the translation.[84] Other targets can be non-coding RNA such as micro-RNA that regulate gene expression.[85,86] A recent study using HTS and sequencing reveals cardiac glycosides as potent inducers of miR-132, a key neuroprotective miRNA downregulated in Alzheimer's disease.[87]

The target protein can also be used as a template for a selection approach that can identify full ligand/inhibitor structures from DEDLs without the need for subsequent fragment linking. The approach of Zhou *et al*. involves dynamic DNA hybridization and target-templated *in situ* ligand synthesis while also incorporates and encodes the linker structures in the library, along with the building blocks, to be sampled by the target protein.[88] They prepared multi-million-member DEDLs with different library architectures and selected hits against four therapeutically relevant target proteins. The dynamic hybridization between encoding DNA attached to building blocks is achieved through a 7 base complementary region, and the target protein shifts the equilibrium towards the production of desired ligands. Only those ligands assembled on the protein are finally decoded and identified. The method's advantage is the reduced need for a high concentration of the target protein and low building block concentration as well as it enables the selection of larger and more versatile chemical libraries.[89]

The first years of this decade have been marked by the COVID-19 pandemic, requiring urgent response from all facets of society, including the medical and scientific community to find new ways of battling its causative agent, the SARS-CoV2 virus.[90] The several billion membered DELs with their capacity for rapid screening have proven a useful tool for detection of compounds of interest in tackling the viral functional components, especially the spike protein and main protease M^{pro}.[91]

One such implementation is a recently reported usage of the RaPID (Random Nonstandard Peptides Integrated Discovery) platform with a genetically encoded library containing constrained macrocyclic peptides for fast identification of several macrocycles with strong binding affinities to the SARS-CoV2 spike glycoprotein and the main protease M^{pro}.[92] Macrocyclic peptides are generally found to be interesting class of potential therapeutics, because they can exhibit strong binding affinities and capability of disrupting their target protein-protein interactions, while also having relatively small MWs, thus being synthetically more accessible than larger proteins (the Goldilocks zone).[93]

This brief overview shows how DELs have advanced the field of drug discovery by providing a practical tool for biological profiling of vast chemical spaces in a test tube (~5 µL per well in microplates). The ability to encode, amplify, and decode complex libraries of compounds has accelerated the discovery of novel bioactive molecules, offering new opportunities for the rapid development of novel therapeutics, especially when medical challenges such as new diseases arise.

In summary, billions of small MW compounds have been synthetized and screened against many therapeutically relevant target macromolecules, mostly proteins, *e.g.* through experiments with various DELs. Structures with possibly physicochemical and biological properties, for millions of synthetic and natural compounds are available through open access online databases including PubChem, ChEMBL, COCONUT *etc*. There are also huge, multi-billion ultra-large screening collections composed of make-on-demand compounds such as ZINC-22, Enamine's REAL, Chemspace's Freedom and and Wuxi's Galaxi®, that can be downloaded to perform VS with the aim of selecting a set of structurally diverse hits that can be synthesised and delivered on request within several weeks.

All these collections, primarily ultra-large combinatorial libraries, also enable fast generation of SARs facilitating the definition of research experiments and projects. This is game-changing as it increases the probability of finding hit molecules while significantly reduces the cost of the early research phase, facilitating the involvement of universities and small and medium-sized enterprises in the discovery of new biologically active molecules.

Acknowledgements. The authors would like to thank the Croatian Ministry of Science and Education for funding this research, and the University of Zagreb University Computing Centre – SRCE for providing advanced computing service.

Data Availability Statement. Data from ChEMBL were used for the visualization analysis.

Conflicts of Interest. The authors declare no conflict of interest.

Abbreviations

ADME(T) = Absorption, Distribution, Metabolism and Excretion (and Toxicology)
B = Billion
BM = Bemis–Murcko
DCL = Dynamic Combinatorial Library
DEDL = DNA Encoded Dynamic Library
DEL = DNA-encoded library
ESAC = Encoded self-assembling chemical library
HBA = a number of hydrogen bond accepting atoms
HBD = a number of hydrogen bond donating atoms
HPC = High-Performance Computing
HTS = High-Throughput Screening
M = Million
ML = Machine Learning
MW = Molecular Weight
NP = Natural Product
NPR = Normalized Principal Moment of Inertia Ratio
PDCL = PNA-Based Dynamic Combinatorial Library
PMI = Principal Moment of Inertia
PNA = Peptide (or polyamide) nucleic acid
QSAR = Quantitative Structure-Activity Relationship
Ro5 = Lipinski's rule-of-five
TC = Tanimoto coefficient
TCM = Traditional Chinese Medicine
VS = Virtual Screening

REFERENCES

[1] J. Quancard, A. Vulpetti, A. Bach, *et al.*, *ChemMedChem* **2023**, *18*, e202300002. https://doi.org/10.1002/cmdc.202300002
[2] F. D. Makurvet, *Med. Drug Discov*. **2021**, *9*, 100075. https://doi.org/10.1016/j.medidd.2020.100075
[3] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50. https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6
[4] C. Knox, M. Wilson, C. M. Klinger, *et al.*, *Nucleic Acids Res*. **2024**, *52*, D1265–D1275. https://doi.org/10.1093/nar/gkad976
[5] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. https://doi.org/10.1016/s0169-409x(00)00129-0
[6] S. Kralj, M. Jukič, U. Bren, *Encyclopedia* **2023**, *3*, 501–511. https://doi.org/10.3390/encyclopedia3020035
[7] J. Carlsson, A. Luttens, *Curr. Opin. Struct. Biol.* **2024**, *87*, 02829. https://doi.org/10.1016/j.sbi.2024.102829
[8] F. Gentile, J. C. Yaacoub, J. Gleave *et al*. *Nat. Protoc.* **2022**, *17*, 672–697. https://doi.org/10.1038/s41596-021-00659-2
[9] T. Zhu, S. Cao, P. C. Su, *et al.*, *J. Med .Chem.* **2013**, *56*, 6560–6572. https://doi.org/10.1021/jm301916b
[10] S. Kim, J. Chen, T. Cheng, *et al.*, *Nucleic Acids Res*. **2023**, *51*, D1373–D1380. https://doi.org/10.1093/nar/gkac956
[11] S. Kim, *Curr. Protoc*. **2021**, *1*, e217. https://doi.org/10.1002/cpz1.217
[12] B. Zdrazil, E. Felix, F. Hunter, *et al.*, *Nucleic Acids Res*. **2024**, *52*, D1180–D1192. https://doi.org/10.1093/nar/gkad1004
[13] J. Hastings, G. Owen, A. Dekker, *et al.*, *Nucleic Acids Res*. **2016**, *44*, D1214–D1219. https://doi.org/10.1093/nar/gkv1031
[14] G. Papadatos, M. Davies, N. Dedman, *et al.*, *Nucleic Acids Res*. **2016**, *44*, D1220–D1228. https://doi.org/10.1093/nar/gkv1253
[15] C. Durinx, J. McEntyre, R. Appel, *et al.*, **2017**, *5(ELIXIR)*, 2422. https://doi.org/10.12688/f1000research.9656.2
[16] M. Sorokina, C. Steinbeck, *J. Cheminform.* **2020**, *12(1)*, 20. https://doi.org/10.1186/s13321-020-00424-9
[17] A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, *et al.*, *Nat. Rev. Drug. Discov*. **2021**, *20*, 200–216. https://doi.org/10.1038/s41573-020-00114-z
[18] M. Sorokina, P. Merseburger, K. Rajan, *et al.*, *J. Cheminform*. **2021**, *13*, 2. https://doi.org/10.1186/s13321-020-00478-9
[19] F. Lovering, *MedChemComm* **2013**, *4*, 515–519. https://doi.org/10.1039/C2MD20347B
[20] A. Rutz, M. Sorokina, J. Galgonek, *et al.*, *Elife*, **2022**, *11*, e70780. https://doi.org/10.7554/eLife.70780
[21] X. Chen, H. Zhou, Y. B. Liu, *et al.*, *Br. J. Pharmacol.* **2006**, *149*, 1092–1103. https://doi.org/10.1038/sj.bjp.0706945

[22] F. M. Afendi, T. Okada, M. Yamazaki, *et al.*, *Plant Cell Physiol*. **2012**, *53*, e1. https://doi.org/10.1093/pcp/pcr165

[23] T. Hoffmann, M. Gastreich, *Drug Discov. Today*. **2019**, *24*, 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013

[24] C. D. Duarte, E. J. Barreiro, C. A. Fraga, *Mini Rev. Med. Chem.* **2007**, *7*, 1108-1119. https://doi.org/10.2174/138955707782331722

[25] E. J. Barreiro, in Privileged Scaffolds in Medicinal Chemistry: Design, Synthesis, Evaluation, ed. S. Bräse, The Royal Society of Chemistry, **2015**, pp. 1–15.

[26] R. B. Merrifield, *J. Am. Chem. Soc.* **1963**, *85*, 2149–2154. https://doi.org/10.1021/ja00897a025

[27] Á. Furka, *Drug Dev. Res*. **1995**, *36*, 1–12. https://doi.org/10.1002/ddr.430360102

[28] F.-M.Klingler, M.Gastreich, O. O.Grygorenko, *et al.*, *Molecules* **2019**, *24*, 3096. https://doi.org/10.3390/molecules24173096

[29] D. F. Veber, S. R. Johnson, H. Y.Cheng, *et al.*, *J. Med. Chem.* **2002**, *45*, 2615–2623. https://doi.org/10.1021/jm020017n

[30] M. J. Waring, J. Arrowsmith, A. R. Leach, *et al.*, *Nat. Rev. Drug. Discov*. **2015**, *14*, 475–486. https://doi.org/10.1038/nrd4609

[31] V. Stepanić, D. Žiher, V. Gabelica-Marković, *et al.*, *Eur. J. Med. Chem*. **2012**, *47*, 462–472. https://doi.org/10.1016/j.ejmech.2011.11.016

[32] C. J. Harris, R. D. Hill, D. W. Sheppard, *et al.*, *Comb. Chem. High Throughput Screen*. **2011**, *14*, 521–531. https://doi.org/10.2174/138620711795767802

[33] M. Congreve, R. Carr, C. Murray, H. Jhoti, *Drug Discov. Today*. **2003**, *8*, 876–877. https://doi.org/10.1016/s1359-6446(03)02831

[34] D. A. Olmedo, A. A .Durant-Archibold, J. L. López-Pérez, J. L. Medina-Franco, *Comb. Chem. High Throughput Screen*. **2024**, *27*, 502–515. https://doi.org/10.2174/1386207326666230705150110

[35] G. W. Bemis, M. A. Murcko, *J. Med. Chem*. **1996**, *39*, 2887–2893. https://doi.org/10.1021/jm9602928

[36] W. H. Sauer, M. K. Schwarz, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987–1003. https://doi.org/10.1021/ci025599w

[37] J. Meyers, M. Carter, N. Y. Mok, N. Brown, *Future Med. Chem*. **2016**, *8*, 1753–1767. https://doi.org/10.4155/fmc-2016-0095

[38] A. Rácz, D. Bajusz, K. Héberger, *J. Cheminform*. **2018**, *10*, 48. https://doi.org/10.1186/s13321-018-0302-y

[39] Marvin was used for generating 2D Bemis–Murcko (BM) scaffolds, Marvin 23.17.0, 2023, ChemAxon (http://www.chemaxon.com).

[40] M. J. Frisch, G. W. Trucks, H. B. Schlegel, *et al.*, Gaussian 16, Revision C.02, Gaussian, Inc., Wallingford CT, **2016**.

[41] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Available online: http://www.R-project.org/

[42] R. Guha, *J. Stat. Softw*. **2007**, *18*, 1. https://doi.org/10.18637/jss.v018.i05

[43] L. J. P. van der Maaten, G. E. Hinton, *J. Mach. Learn. Res*. **2008**, *9*, 2579.

[44] L. McInnes, J. Healy, J. Melville, **2018**, *arXiv:1802.03426*. https://doi.org/10.48550/arXiv.1802.03426

[45] J. H. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. **2015**. R package version 0.17, https://github.com/jkrijthe/Rtsne

[46] T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model*. **2015**, *55*, 460–473. https://doi.org/https://doi.org/10.1021/ci500588j

[47] V. Stepanić, M. Kučerová-Chlupáčová, *Molecules* **2023**, *28*, 475. https://doi.org/10.3390/molecules28020475

[48] W. A. Warr, M. C. Nicklaus, C. A. Nicolaou, M. Rarey, *J. Chem. Inf. Model*. **2022**, *62*, 2021–2034. https://doi.org/10.1021/acs.jcim.2c00224

[49] J. Lyu, S. Wang, T. E. Balius, *et al.*, *Nature*, **2019**, *566*, 224–229. https://doi.org/10.1038/s41586-019-0917-9

[50] J. J. Irwin, K. G. Tang, J. Young, *et al.*, *J. Chem. Inf. Model.* **2020**, *60*, 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675

[51] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model*. **2005**, *45*, 177–182. https://doi.org/10.1021/ci049714+

[52] B. I. Tingle, K. G. Tang, M.Castanon, *et al.*, *J. Chem. Inf. Model*. **2023**, *63*, 1166–1176. https://doi.org/10.1021/acs.jcim.2c01253

[53] O. O. Grygorenko, D. S. Radchenko, I. Dziuba, *et al.*, *iScience*, **2020**, *23*, 101681. https://doi.org/10.1016/j.isci.2020.101681

[54] InfiniSee version 6.1.1; BioSolveIT GmbH, Sankt Augustin, Germany, **2024**, https://www.biosolveit.de/infiniSee

[55] M. V. Protopopov, V. V. Tararina, F. Bonachera, *et al.*, *Mol. Inform*. **2024**, e202400114. https://doi.org/10.1002/minf.202400114

[56] R. Kiss, M. Sandor, F. A. Szalai, *J. Cheminform.* **2012**, *4*, 1. https://doi.org/10.1186/1758-2946-4-S1-P17

[57] G. Takács, D. Havasi, M. Sándor, *et al.*, *ACS Med. Chem. Lett*. **2023**, *14*, 1188–1197. https://doi.org/10.1021/acsmedchemlett.3c00146

[58] J. Sunseri, D. R. Koes, *Nucleic Acids Res*. **2016**, *44*, W442–W448. https://doi.org/10.1093/nar/gkw287

[59] A. Brenner, R. A. Lerner, *Proc. Natl. Acad. Sci*. **1992**, *86*, 5381–5383.

[60] J. K .Scott, G. P. Smith, *Science*. **1990**, *249*, 386–390. https://doi.org/10.1126/science.1696028
[61] A. D. Ellington, J. W. Szostak, *Nature*, **1990**, *346*, 818–822. https://doi.org/10.1038/346818a0
[62] A. Furka, F. Sebestyén, M. Asgedom, G. Dibó, *Int. J. Pept. Protein Res.* **1991**, *37*, 487–493. https://doi.org/10.1111/j.1399-3011.1991.tb00765.x
[63] L. A. Thompson, J. A. Ellman, *Chem. Rev.* **1996**, *96*, 555–600. https://doi.org/10.1021/cr9402081
[64] M. C. Needels, D. G. Jones, E. H. Tate, *et al.*, *Proc. Natl. Acad. Sci. U S A.* **1993**, *90*, 10700–10704. https://doi.org/10.1073/pnas.90.22.10700
[65] S. Melkko, J. Scheuermann, C. E. Dumelin, D. Neri, *Nat. Biotechnol*. **2004**, *22*, 568–574. https://doi.org/10.1038/nbt961
[66] S. Oehler, L. Plais, G. Bassi, *et al.*, *Chem. Commun.* **2021**, *57*, 12289–12292. https://doi.org/10.1039/d1cc04306d
[67] M. Cui, D. Nguyen, M. P. Gaillez, *et al.*, *Nat. Commun*. **2023**, *14*, 1481. https://doi.org/10.1038/s41467-023-37071-1
[68] L. Plais, A. Lessing, M. Keller, *et al.*, *Chem. Sci*. **2022**, *13,* 967–974. https://doi.org/10.1039/d1sc05721a
[69] D. R. Halpin, P. B.Harbury, *PLoS Biol*. **2004**, 2, E174. https://doi.org/10.1371/journal.pbio.0020174
[70] M. A. Clark, R. A. Acharya, C. C. Arico-Muendel, *et al.*, *Nat. Chem. Biol.* **2009**, *5*, 647. https://doi.org/10.1038/nchembio
[71] I. Huc, J. M. Lehn, *Proc. Natl. Acad. Sci U S A*. **1997**, *94*, 2106–2110. https://doi.org/10.1073/pnas.94.6.2106
[72] S. Otto, R. L. Furlan, J. K. Sanders, *Drug Discov. Today*, **2002**, 7, 117–125. https://doi.org/10.1016/s1359-6446(01)02086-4
[73] P. A Brady, R. P. Bonar-Law, S. J. Rowan, *et al.*, *Chem. Commun*. **1996**, 319–320. https://doi.org/10.1039/CC9960000319
[74] S. Ladame, *Org. Biomol. Chem*. **2008**, *6*, 219–226. https://doi.org/10.1039/b714599c
[75] F. Cardullo, M. C. Calama, B. H. M. Snellink-Ruël, *et al.*, *Chem. Commun*. **2000**, 367–368. https://doi.org/10.1039/A909459H
[76] K. C. Nicolaou, R. Hughes, S. Y. Cho, *et al.*, *Chemistry*. **2001**, *7*, 3824–3843. https://doi.org/10.1002/1521-3765(20010903)7:17<3824::aid-chem3824>3.0.co;2-1
[77] W. G. Lewis, L. G. Green, F. Grynszpan, *et al.*, *Angew. Chem. Int. Ed. Engl*. **2002**, *41*, 1053–1057. https://doi.org/10.1002/1521-3773(20020315)41:6<1053::aid-anie1053>3.0.co;2-4
[78] D. A. Erlanson, J. W. Lam, C. Wiesmann, *et al.*, *Nat. Biotechnol*. **2003**, *21*, 308–314. https://doi.org/10.1038/nbt786
[79] M. Dockerill, N. Winssinger, *Angew. Chem. Int. Ed. Engl*. **2023**, *62*, e202215542. https://doi.org/10.1002/anie.202215542
[80] P. E. Nielsen, M. Egholm, R. H. Berg, O. Buchardt, *Science*. **1991**, *254*, 1497–1500. https://doi.org/10.1126/science.1962210
[81] M. Egholm, O. Buchardt, L. Christensen, *et al.*, *Nature*, **1993**, *365*, 566–568. https://doi.org/10.1038/365566a0
[82] L. Farrera-Soler, J. P. Daguer, P. Raunft, *et al.*, *Bioorg. Med. Chem.* **2020**, *28*, 115458. https://doi.org/10.1016/j.bmc.2020.115458
[83] K. D. Warner, C. E. Hajdin, K. M. Weeks, *Nat. Rev. Drug Discov*. **2018**, *17*, 547–558. https://doi.org/10.1038/nrd.2018.93
[84] C. J. Radoux, F. Vianello, J. McGreig, N. Desai, A. R. Bradley, *Front. Bioinform*. **2022**, *2*, 958378. https://doi.org/10.3389/fbinf.2022.958378
[85] S. P. Velagapudi, S. M. Gallo, M. D. Disney, *Nat. Chem. Biol.* **2014**, *10*, 291–297. https://doi.org/10.1038/nchembio.1452
[86] S. P. Velagapudi, M. D. Cameron, C. L. Haga, *et al.*, *Proc. Natl. Acad. Sci. U S A*. **2016**, *113*, 5898–5903. https://doi.org/10.1073/pnas.1523975113
[87] L. D. Nguyen, Z. Wei, M. C. Silva, *et al.*, *Nat. Commun*. **2023**, *14*, 7575. https://doi.org/10.1038/s41467-023-43293-0
[88] Y. Zhou, W. Shen, Y. Gao, *et al.*, *Nat. Chem*. **2024**, *16*, 543–555. https://doi.org/10.1038/s41557-024-01442-y
[89] D. Bosc, J. Jakhlal, B. Deprez, R. Deprez-Poulain, *Future Med. Chem*. **2016**, *8*, 381–404. https://doi.org/10.4155/fmc-2015-0007
[90] T. Asselah, D. Durantel, E. Pasmant, G. Lau, R. F. Schinazi, *J. Hepatol*. **2021**, *74*, 168–184. https://doi.org/10.1016/j.jhep.2020.09.031
[91] R. Jimmidi, S. Chamakuri, S. Lu, *et al.*, *Commun. Chem*. **2023**, *6*, 164. https://doi.org/10.1038/s42004-023-00961-y
[92] S. Ullrich, C. Nitsche, *Isr. J. Chem.* **2024**, online e202300170, https://doi.org/10.1002/ijch.202300170
[93] C.Morrison, *Nat. Rev. Drug Discov*. 2018, 17, 531–533. https://doi.org/10.1038/nrd.2018.125