Molecular Similarity: Theory, Applications, and Perspectives

Kenneth López-Pérez,[1] Juan F. Avellaneda-Tamayo,[2] Lexin Chen,[1] Edgar López-López,[2,3] K. Eurídice Juárez-Mercado,[2] José L. Medina-Franco,[2*] Ramón Alain Miranda-Quintana[1*]

[1] *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida, 32611, USA*

[2] *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*

[3] *Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Section 14-740, Mexico City 07000, Mexico*

* Emails: quintana@chem.ufl.edu, medinajl@unam.mx

Abstract

Molecular similarity pervades much of our understanding and rationalization of chemistry. This has become particularly evident in the current data-intensive era of chemical research, with similarity measures serving as the backbone of many Machine Learning (ML) supervised and unsupervised procedures. Here, we present a discussion on the role of molecular similarity in drug design, chemical space exploration, chemical "art" generation, molecular representations, and many more. We also discuss more recent topics in molecular similarity, like the ability to efficiently compare large molecular libraries.

Introduction

Similarity is essential to human cognition because it enables us to generalize characteristics along a category or to classify items in the universe according to an ordered array of sets whenever they share a particular feature ref.[1–4] The possibility of inferring some knowledge about a presumptive shared property between two similar items depends on the mentioned attribute and the preexistent relationship between it and the shared properties that make the two objects similar.[5] Depending on the particular character studied, similarity is a subjective reflection of studied objects.[6] These postulates are highly related to the *similarity-property principle* widely applied in medicinal chemistry, as much as in other scientific areas which enunciates that "similar structural features give rise to similar properties/biological activity". However, generalizing shared properties has been employed in chemical sciences centuries before a consensus on atomic structure.[7] For instance, to approximate chemical reactivity and physical properties as criteria that permit categorizing elements or understanding the matter structure. Whether in Lavoisier's attempts to complete the rules for nomenclature in chemistry, or classify bodies by their elements and compounds they decompose to, and the categorization of those elementary components between metals and nonmetals[8]; or the reach for associative relationships

of triads, by Döbereiner in a sort of a systematic classification of “analogous” elements, deducting relationships on atomic masses preserved in the current periodic table.[9] Eforces would derive in the current robust but dynamic table of elements, with twice as many elements as at the time of its postulation one hundred fifty years ago by Mendeleev. A remarkable fact from this background is the precise prediction of the chemical and physical properties of unknown elements for his time, which would be discovered in the next three decades, confirming his suppositions.[10]

Similarity is a concept applied systematically in the exercise and mission of science to explain and categorize phenomena and bodies by identifying relations among a population. This definition is applicable in fields as diverse as biology taxonomy, which is the methodology and principles of systematic arrangements for living and extinct organisms[11]; etymology, which studies the origin and relationships among languages through the similarity and root of written and spoken words, or software sciences, where classification of structures and architectures allows a better relationship between academic and industrial applications of software development.[12]

In Chemistry, similarity refers to the common functionalities, structures, composition, spatial disposition, biological activity, and physicochemical properties among different chemical compounds, biological systems, and macromolecular complexes, among others. Similarity has become a cornerstone of chemoinformatics[6], which makes it of great interest to chemists and pharmaceutics, as well as appearing in other diverse domains (see the following sections). Based on the *similarity-property principle*, similarity becomes one of the main approaches to assessing problems related to molecular screening in different fields with chemical applications, giving rise to molecular similarity analysis. It has been largely demonstrated through the practice that results in similarity, and so on the similarity concept *per se* is subjective in the sense that different representations used in the chemical description and different approximations to similarity quantitation bring different criteria of similarity.[6,13]

1. Molecular Similarity

1.1 Quantification of molecular or system similarity

How to describe a molecule? How to describe a chemical-biology system, e.g., a small molecule bound to a protein? These are interesting and philosophical questions with multiple correct answers and have significant

implications for molecular design. To address these questions, firstly, it is essential to answer “What are the molecular properties that we want to explore, study, or decode?” Namely, many descriptors and chemical representations exist that illustrate specific parts of a molecular dataset's most complex property, e.g., bioactivity, that depends on structural, physicochemical, and metabolic features, etc. Molecular representation and description are at the cornerstone of basically any computational method. The representation and descriptors choice will depend on a number of major factors including, but not limited to, the study’s goals; type of compounds under study e.g., small organic molecules, peptides, flexible macrocycles, inorganic and metal-containing molecules, etc.), and number of compounds.[14]

Classic and non-classic representations have been extensively used to decode different molecular features like fingerprints-, scaffolds-, and graph-based representations that have key differences, and each one is useful to represent different features of a molecular dataset. For example, MACCS Keys is one of the most used dictionary-based molecular fingerprints that represent the presence and absence of key functional groups; Extended Connectivity Fingerprints (ECFPs) allow for representing the spatial distribution and connectivity of each atom in each molecule in a dataset (see Table 1.1_1); Graph-based fingerprints that allow condense different kind of molecular information in a unique matrix for each molecule in a dataset like structural, stereochemical, atomistic, and three-dimensional data.

Table 1.1_1. Classical molecular representations and descriptors: standard and non-standard. Adapted from (Martinez-Mayorga Karina, et al. The pursuit of accurate predictive models, 2023).

Endpoint described	Classic descriptors	Non-classic descriptors
General structure-property relationships	• Classical molecular fingerprints e.g., MACC keys, PubChem, ECFP. • Chemical diversity descriptors e.g. functional groups and Bemis-Murko scaffolds.	• Non-classical (and recently developed) molecular fingerprints (e.g. MAP4, and atom pairs). • Graph-based representations. • Sequence-based representations. • Spectra-based representations.
Specific industrial applicabilities	• Druglike properties e.g. LogP, molecular weight.	• Organoleptic properties (e.g. odor or flavor). • Material properties (e.g. conductivity).
ADMET predictions	• Qualitative ADMET descriptors e.g., Inhibitor of cytochromes.	• Quantitative ADMET descriptors (e.g. clearance, bioavailability, half-life time).
Reactivity	• Polarizability.	• Quantum descriptors
Biological and bioactive	• Bioactivity e.g., enzymatic or cell-grown inhibition. • Phenotypic effects.	• Post-marketing data (e.g. drug safety in different populations) • -OMICS data (e.g. pharmacogenomic or proteomic data).

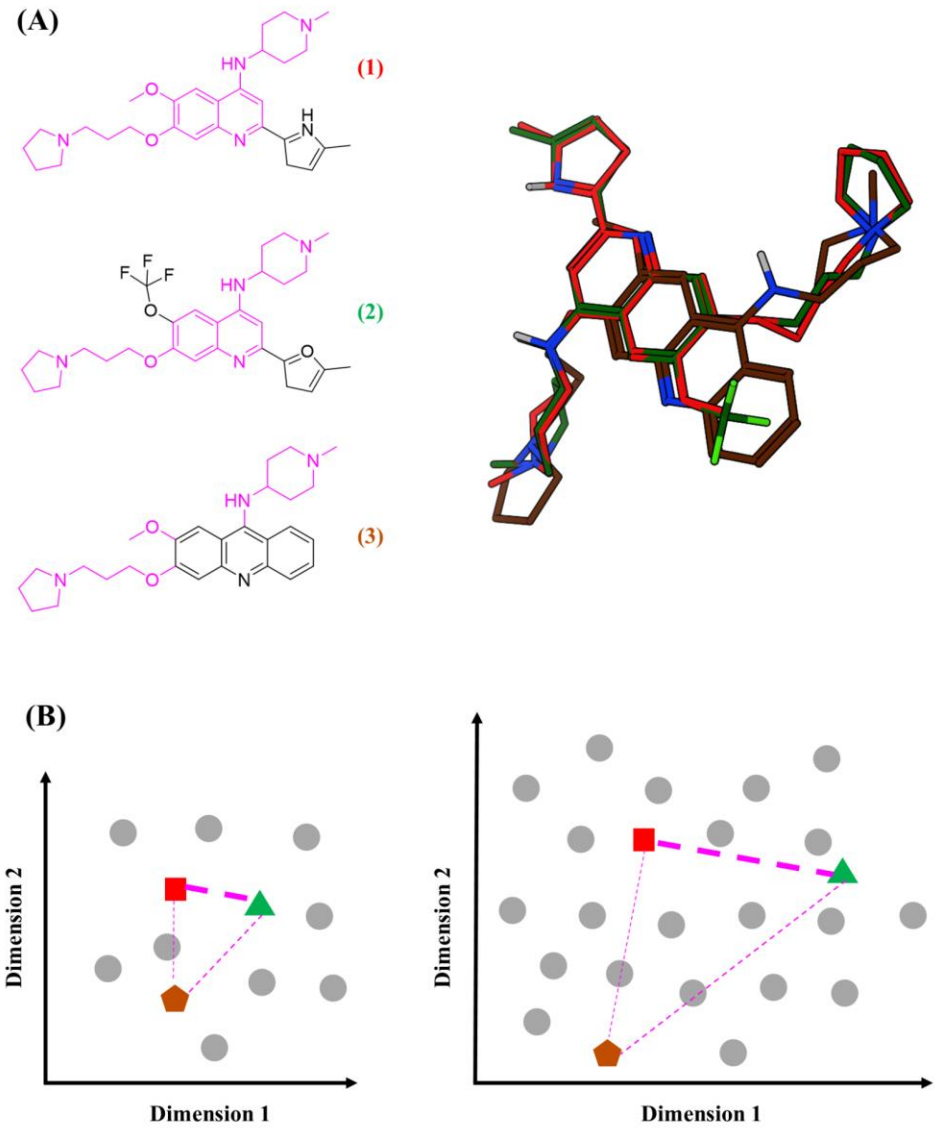
In parallel, there are classical and non-classical descriptors like drug-likeness, ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxic) properties, reactivity, and biological features (Table X1). That has impacted in important molecular design areas, for example, to decode structure-property relationships (SPRs), reactivity, and biological issues. For example, drug-likeness and ADMET descriptors have been broadly used to improve the potency, selectivity, and safety of drugs; Reactivity descriptors have been used to describe, compare, and predict molecular mechanisms, chemical reactions, and most stable conformational states; Biological and bioactive descriptors, e.g., phenotypic, post-marketing, or -omical data, that allows decoding physiological effects of molecules against complex biological systems.

Peptides are interesting molecules or "compound systems" that, mostly depending on their size (few to several and large sequences) are at the interface between representations typically used in chemoinformatics (if short amino acids with few amino acids are regarded as small molecules) or bioinformatics (like sequence-based representations). However, there is not a unique opinion about the number of amino acids (molecular weight, or the number of atoms) that must be considered to use chemoinformatics or bioinformatics representations. Recently, different authors suggested that peptides lower than 40 amino acids must be represented using chemoinformatics approaches, and the higher than must be represented using bioinformatics approaches.[15,16] However, the authors remark on the importance of establishing a clear main objective prior to selecting the descriptors used to represent any molecule.

1.2 The relative and subjective nature of similarity

Representation of a small molecule or molecular system, plus a similarity metric, is essential to quantify similarity between a pair of systems. Since the "best" or more appropriate representation depends on the goals of the comparison itself, for example, the goal of the project, similarity is strongly subjective. It follows that the real similarity between any pair of objects/systems is not absolute and it is context-dependent (where the "context" not only is associated with the representation or descriptors but also with the "environment" or neighbors".[17] Perhaps one of the most straightforward manners to exemplify the neighborhood dependence of similarity is a two-dimensional (2D) vs. three-dimensional (3D) similarity assessment. The latter 3D-

similarity also known as 3D superposition has been recently reviewed.[18] The concept of context-dependent similarity is outlined in Figure 1_2A.



[Figure 1_2]. **A**) Context-dependent system/molecule similarity. Exemplary comparisons of three small molecules highlighting their similarity in substructures and scaffolds (marked in pink); and their similarity in three dimensions. Adapted from.[19] **B**) Time-dependent similarity: two objects (systems, molecules) can change similarity over time, depending on the modifications or changes in the molecular descriptors (lower vs. higher dimensionality). The figure illustrates the hypothetical similarity and distance between the compounds being compared. The thickness of the pink lines represents the hypothetical structural similarity values.

It also follows that the similarity between two systems is not necessarily "absolute" in terms of time but it can vary with time (Figure 1_2B). In other words, similarity can be time-dependent. Of course, the time dependence will be significant if the descriptors change with time. Similar to the goals of any comparison,

quantifying the evolution with time of the similarity will depend on the project's goals. The next section describes the most common molecule and system representations.

1.3 Representation

Molecular representation has been of common interest for research in chemistry for at least two centuries. Generally, this is achieved by using diagrams of element symbols (or vertex of a graph) and sticks (edges) that represent bonds. This way of representation has also been an attempt to explain reality, and how atoms join through electronic interactions to form molecules, as planted by Lewis even before a consensus in the structure of the atoms.[20] Also, a condensed form to denote the composition of chemical substances, both organic and inorganic, has been accomplished by omitting certain characteristics in connectivity and geometry.[21]

The same procedure has been used in the representation of biomolecules such as peptides or proteins, and nucleic acids, by representing their whole building blocks with a system of well-established symbols of three letters for amino acids, and one letter for amino acids and nitrogen bases. In biochemical systems with the participation of covalent, and moreover non-covalent interactions, the sequence of building blocks does not explain these interactions. Then, as complexity increases, representation has to be more complex to achieve a suitable coverage of structural, electronic, and physicochemical features, which is a key question in bioinformatics.[21]

For organometallic structures, electronic singularities difficult the issue of representation. Whereas in the case of augmented valances, chirality around metallic centers, or the limited handling of molecular geometry, that is why there are specific representations for this kind of system.[22]

However, the computational approach requires the implementation of a systematic method of input, processing, and storage of chemical structures, that cover a different degree of chemical properties, e.g. connectivity, shape, charges, and physicochemical properties, among others. This approach is accomplished by the representation of molecules through molecular graphs and their posterior transformation by the implementation of different formats into strings and matrices readable by computers. In accomplishing the goal of an adequate chemical representation, adjusted to a specific research objective, many useful tools have been developed, some of them more specific or, on the contrary, versatile, but with a plethora of different

application areas. The issue of representation has been widely discussed, and recently reviewed in its application to different fields.[21,23,24]

Over the last decades, with the technological advances in computation, and the development of new methodologies and tools, there has been a rapid emergence of different and innovative ways to represent chemical compounds, from small organic molecules to macromolecules, and chemical reactions.[21] Some of them are revisited in the next subsections.

1.3.1 Linear notations

Linear notations are one of the most used representations. Those are cheap in computational resources and space, and with time have been standardized and widely used, which is the case for Simplified Molecular Input Line Entry Systems (SMILES) or International Chemistry Identifiers (InChI). They store information in the form of a 1D letter string, where most of the characteristics of the molecule are implicit by facility, such as geometry, hydrogen atoms linked to the main chain, and aromaticity. However, this information can be extracted by following the precise rules associated with the notation.[25] SMILES have been the most accepted notation because of its human readability, despite not being singular for a molecule, it can be easily converted into a canonical version by implementing consensual rules available in different toolkits such as RDKit, implemented in the Python programming language.[26]

InChI notation has emerged as a solution to the problem of unicity in 1D representation, open source accessibility, and a big data management solution.[27] As SMILES strings, they encode chemical structures as 1D strings, specifying in their first block the molecular skeleton and isomerism in the second. InChI notation is not as interpretable by intuition as SMILES but is also required to SMILES canonicalization.[28] SMILES Arbitrary Target Specification (SMARTS) has been developed for partial terms of searching, for which it is possible to use generic parts of the string to specify a changeable fragment of the molecules.[29]

1.3.2 Molecular fingerprints

Molecular fingerprints consist of the description of properties or substructures present in the molecules, either by the focused search of specific and predefined features (dictionary-based fingerprints) or by an independent and mathematical description of the molecular characteristics (circular, topological, pharmacophore, protein-

ligand interaction, shape-based, reinforced, and multi fingerprints). This respect has been recently reviewed and robustly classified.[30] Chemical fingerprints were developed to relate molecular structural features and physicochemical properties, by their inclusion into QSAR equations[31], and have evolved into one of the most systematic and broadly used methodologies for molecular representation in different computational applications.[32] Molecular features and chemical descriptors are included in chemical fingerprints as components of a vector or tuple which can be handled by computers, and permit the realization of mathematical transformations.[33] Below are some of the most important and widely used chemical fingerprint categories and examples of their specific architectures.

Dictionary-based structural fingerprints, keys fingerprints, or molecular independent fingerprints, are specified sets of features examined for in molecules, and their presence (1) or absence (0) is recorded as a bit vector position. These features can be functional groups, characteristic substructures, or fragments ref. The most representative dictionary-based structural fingerprints are PubChem (PC)[34], Molecular ACCess System (MACCS)[35], Mini FingerPrint (MFP)[36], Barnard Chemistry Information (BCI) fingerprints[37], and SMIles FingerPrint (SMIFP).[38]

All the other types of fingerprints consist of dependent architectures. Those descriptors have a variable number of components, and their design depends on particular connectivity features of studied molecules, which makes them more specific and appropriate for complex chemical structures.

Circular fingerprints are those in which information capturing is made from a central heavy atom and explores the surrounding neighbors iteratively, until a complete description of the structural features.[39] Examples of circular fingerprints are Extended Connectivity FingerPrints (ECFPs)[40], Functional-Class FingerPrints (FCFPs)[41], Molprint2D[42], and Molprint3D[43].

Topological fingerprints are those that represent the connectivity of the molecules in terms of nodes (atoms) and edges (bonds).[44] This kind of description gives rise to a connection table such as SDF and MOL format, that specifies the nature of the atom, connectivity, bond distances, atomic eccentricity, and/or weight. Examples of topological fingerprints are atom pairs fingerprints (APs)[45], topological torsion (TT)[46], and Daylight fingerprints[47].

There are other types of fingerprints aimed at studying molecular systems that involve the interaction between two molecules, for instance, a protein-ligand complex. Examples are pharmacophore fingerprints

and protein-ligand interaction fingerprints. Pharmacophore fingerprints describe the necessary characteristics of the ligand according to their possible interactions, and three-dimensional characteristics such as hydrogen bonding, charge transfer, electrostatic, and hydrophobic interaction sites. Protein-ligand interaction fingerprints (PLIF) collect information on explicit interactions between protein and ligands, by describing the bonding amino acid residues present in the protein site, and geometric disposition.[48] Such 3D information is provided by molecular docking or experimental results, which are transformed into 1D bitstrings. Broadly used examples are structural interaction fingerprints (SPLIFs)[49], and protein-ligand extended connectivity (PLEC) fingerprints.[50]

Shape-based fingerprints are more recently developed and consist of a description in terms of the probable 3D interacting surface of the molecule, through the collection of coordinates, and potential interacting types. Some shape-based fingerprints of broad use are rapid overlay of chemical structures (ROCS) and ultrafast shape recognition (USR).[51]

1.3.3 Molecular properties

As structural features, physicochemical and constitutional properties can be used to describe and represent molecules and sets of molecules, using continuous and/or discrete features, due to their intrinsic relationship with structural characteristics. These features are known as classical molecular descriptors.[52] Their utility in computational chemistry and chemoinformatics has been widely reviewed recently.[53,54] Among most classical and relevant descriptors there are tendencies calculated and studied principally in drug discovery, that empirically showed a range of acceptability. Those properties are part of empirical rules to quantify drug-likeness such as the rule of five by Lipinski[55], and its extension by Veber.[56] Typical properties to quantify drug-likeness include molecular weight, number of carbon atoms, number of oxygen atoms, ring systems, count of single or multiple bonds, number of hydrogen bond donor/acceptors, and the logarithm of the octanol-water partition coefficient, logP.

1.3.4 Chemical reactions

Transformation of a chemical compound into another following logical and chemistry-sensed rules of transformation is of central interest in chemoinformatics and computational chemistry, due to their high

prevalence (more than 150 million records at the moment of writing)[57] and common study, in predictive models, searching, storing, and categorizing chemical information[58,59]. Common elements of a reaction such as reactants, conditions, and products, must be included in a computer-readable format. Some of the most widely employed representations of chemical reactions are Reaction SMILES, SMIRKS, and RInChI, among others.

Reaction SMILES are composed of the representation of reactants, agents, and products, as SMILES notation, and the reaction arrow symbolized by >. Atoms in molecular representations can be ticked to follow a specific pattern of displacement along a chemical reaction. This is achieved by the implementation of numerical indices.[60] SMIRKS notation is the SMARTS analogous to Reaction SMILES, which can represent generic reaction patterns. Just exchangeable groups or substructures have to be explicitly represented by SMIRKS, and the rest of the molecule can be codified by generic symbols. SMARTS are used to represent unchangeable generic substructures.[29,61]

RInChI uses InChIs to represent molecules implicated in the reaction, and its layered structure includes information about reaction conditions, catalyst, and solvent. As the InChI and InChI key, RInChI provides a standardized and non-ambiguous way to represent chemical reactions and is very powerful for data management.[62] Additional representations of chemical reactions are condensed graphs of reaction (CGR)[63] and bond electron matrices (BE-matrix).[64,65]

1.3.5 System representations

The most explored chemical representations have been applied to small molecules, however, there has been a necessity to develop suitable depictions for the expression and operation with macromolecules and more complex systems. Chemoinformatics and bioinformatics have contributed to this endeavor.

The one-letter abbreviation has been the most common and abbreviated linear notation for macromolecules such as peptides, proteins, and nucleic acids. However, as mentioned before, there are some limitations such as the limited characters of the Latin alphabet, or the lack of structural information using them, in particular, 3D information.

CHUCKLES and CHORTLES are classical examples of the representation of macromolecules at sequential and atomic levels. CHUCKLES consist of an interpreter that translates the peptidic sequence into

the molecular complex SMILES, including covalent interactions along the chain.[66] Equivalently, CHORTLES is capable of handling oligomeric mixtures.[67] In both cases, cyclization, branching, or mixtures are expressed with special characters. In turn, the self-contained sequence representation (SCSR), uses the V3000 molfile format to record biopolymer structures in a compressed format, keeping structural details, and remaining efficient for information management .[68] A more recent development in the chemical representation of proteins and biomolecules is the Hierarchical Editing Language for Macromolecules, or HELM notation language, which is a code at the polymeric level (SMILES does this at the atomic level), giving rise to a standardized and unified representation of biomolecules for both bioinformaticians and chemoinformaticians.[69]

Specific examples of macromolecular representations are the web3 unique representation of carbohydrate structures (WURCS), for oligosaccharides and polysaccharides with three to twenty monosaccharides of interest because of their role in biochemical interactions and molecular recognition. WURCS can encode for the main carbon backbone, and also for modifications and branches in glycans.[70,71] BigSMILES is an emergent canonical representation for macromolecules such as biopolymers, based on SMILES linear notation. BigSMILES has started to be widely implemented in applications such as big data management in the study of biopolymers.

2. Applications, practical implications, and implementations

Figure [2_1] summarizes schematically various practical applications of the similarity concept. Although the applications can be organized in various different but valid classifications, the applications can be divided depending on the research area, namely, chemistry, biology, and clinical applications. The approaches can also be organized into the following categories in general: similarity searching, database mining, and compound datasets (chemical libraries) profiling that could include the profiling or analysis of structure-property relationships that, as shown in this Section, have strong implications in medicinal chemistry.[17]

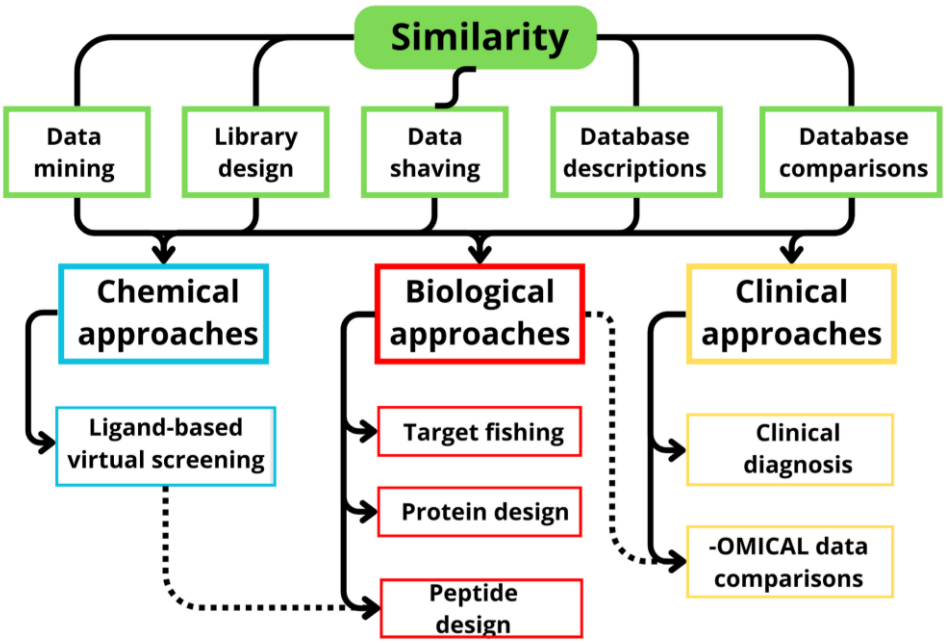


Figure [2_1]. Overview of different applications and implications of similarity. The solid lines represent the relationships between the widely documented areas/approaches, while the dashed lines symbolize emergent relationships.

In many practical applications, molecule (chemical) or system similarity is extensively used in similarity searching that is founded on the "similarity principle"[72] if two molecules have similar structures it is anticipated that they will have similar properties. In drug discovery, the property is biological activity so the principle can be formulated as "similar compounds have similar (biological) activity". However, well-known exceptions of such principle are the property cliffs: defined as pairs of compounds with high structure similarity but very large (and unexpectedly) different properties.[73] The equivalent in drug discovery, "activity cliffs" proposed in 2006 by Maggiora[74] has been the subject of extensive research in terms of elucidation, implications in drug discovery, and influence while trying to develop predictive models. Property and activity cliffs, which arguably can be artifacts of the experimental assessment to measure the relevant property[75], have been the subject of several review papers, and the reader is referred to such publications.[76,77] Similarity searching is extensively applied in ligand-based virtual screening (comparison of one or more reference or query compounds vs., a set of molecules in a chemical library; in target fishing.

One of the implementations of similarity searching is in websites to perform database mining. Many web-based compound databases in the public domain have implemented a "similarity search" function where the user has the option to select or fix a given query chemical structure and the server/website will provide a set of "similar compounds" based on a threshold defined by the user (many websites have a filer so the user can

set up a threshold value of similarity). One of the downsides of such databases is that not always is clear the internal representation used by the server (a crucial point during the quantification of similarity (Section 1). Table [2_1] summarizes examples of websites that have an implementation of similarity searching. The table presents just representative applications to conduct similarity searches on small molecules and peptides, and spectra.

Table [2_1]. Exemplary websites with the implementation of similarity searches.

Website name	Main application	similarity function	Ref. / URL
ChemSpider	Small molecule and peptide similarity searches.	Structure, substructure, and scaffolds similarity searches	http://www.chemspider.com/
DrugBank			https://go.drugbank.com/
ZINC			https://zinc.docking.org/
Sci-Finder		Structure, substructure, scaffold, and connectivity similarity searches	https://scifinder.cas.org
ChEMBL			https://www.ebi.ac.uk/chembl/
COCONUT		Identity, structure, substructure, scaffold, and connectivity similarity searches. Searches by a range of continuous properties.	https://coconut.naturalproducts.net/
NuBBE			https://nubbe.iq.unesp.br/portal/nubbe-search.html
UnitProt	Peptide and protein similarity searches	Sequence and domain similarity searches	https://www.uniprot.org/
BLAST - NCBI			https://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA			https://www.ebi.ac.uk/Tools/sss/fasta/
Protein Data Bank		Sequence, structure, and chemical similarity searches	https://www.rcsb.org/
CSEARCH	Searches based on ^{13}C RMN spectra	Signals, elements, and weight similarity searches	https://c13nmr.at/similar/eval.php
CFM-ID	Compounds identification based on MS spectra similarity	Signal similarity searches	https://cfmid.wishartlab.com/

Another major implication of molecular similarity is profiling or characterization of similarity of compound data sets and diversity analysis. Similarity or diversity profiling of compound data sets is very useful in different

areas such as library selection and design including *de novo* design of chemical libraries; comparison of similarity profiles with implications, for instance, in natural product research (with implications in ecology: assessment of the biodiversity of ecosystems).

In general, since there are many alternatives to measure similarity (or distance) (*vide supra*), Miranda-Quintana et al. have proposed the Differential Consistency Analysis. It was found that the consensus between Tanimoto and the Cosine coefficients improved by choosing a query whose similarity to the rest of the molecules varies less, or by describing the molecules in a manner that does not depend strongly on their size.[3]

2.1 In drug design and discovery

Similarity has an important role in the realm of drug discovery design, where the quest for novel therapeutic compounds often depends on identifying molecules that exhibit structural and functional similarities to known drugs or bioactive compounds. Levering similarities metrics, such as molecular fingerprinting or ligand-based approaches, by recognizing similarities in chemical structures or biological activities the process of drug development could be significantly shortened.[78] This approach highlights the crucial role of similarity-based strategies as essential tools in the pursuit of identifying new compound candidates to be novel drugs.

Table [2.1_1] provides a comprehensive overview of the diverse applications of similarity in the field of drug design. These applications harness the power of chemical and structural similarity to expedite the discovery, design, and optimization of drug compounds. From virtual screening to quantitative structure-activity relationship (QSAR) modeling and scaffold hopping, each application is important in different stages of drug development. The table shows representative examples and recent advances, showcasing how computational techniques and chemoinformatics continue to play a significant role in revolutionizing the drug discovery process, ultimately leading to novel therapeutic compounds and more efficient drug development pipelines.

Table [2.1_1]. Selected applications of similarity in drug design.

Application	Description	Methods and tools used	Key advantages	Representative examples	Ref.

Virtual screening	Identifying potential drug candidates by comparing the chemical similarity of molecules to known active compounds.	Molecular docking.	Efficient screening of large chemical libraries.	Screening a chemical library against a target protein using Tanimoto coefficient.	79,80
		Ligand-based methods.	Reduces experimental work.		
		Chemoinformatics tools like RDKit.			
Drug repurposing	Discovering new therapeutic uses for existing drugs by identifying similarities between the target and known drug-protein interactions.	Data mining of chemical genomic databases.	Accelerates drug discovery process.	Repurposing antiviral drugs for cancer treatment based on shared molecular pathways.	81,82
		Computational network analysis.	Lower risk and cost compared to *de novo* drug development.		
		Machine learning.	Potential for quick clinical translation.		
Quantitative-Structure Activity Relationship (QSAR)	Predicting the biological activity or toxicity of compounds based on their chemical structures and similarity known data.	Molecular modeling.	Enables rational drug design.	Building QSAR models using chemical descriptors and similarity measures to predict drug interactions or toxicity profile.	31,83
		Machine learning algorithms.	Reduces the need for extensive experimental testing.		
		Descriptor calculation tools.	Enhances compound optimization.		
Scaffold hopping	Designing new drug molecules by replacing or modifying specific structural components of existing compounds while maintaining desired properties.	Scaffold similarity metrics	Facilitates the discovery of novel chemical compounds with known pharmacological profiles.	Designing new molecules with improved properties based on existing scaffolds.	84,85
		Scaffold analysis.	Enhances compound diversity in libraries.		
Chemogenomics	Integrating chemical and genomic data to identify relationships between compounds and	Chemical genomic data integration.	Enables the discovery of new drug-target interactions.	Analyzing chemical genomic datasets to find links between compounds and biological targets.	86,87
		Network analysis.	Provides insights into compound mechanisms of		

	target proteins or pathways.		action.		
		Machine learning.	Facilitates polypharmacology studies.		
3D Structural similarity	Comparing the three dimensional structures of molecules to identify potential binding sites and interactions with target proteins.	Molecular docking simulations.	Offers insights into binding models and interactions.	Molecular docking to assess structural similarity between ligands and binding sites.	88,89
		Structural alignment tools.	Enables rational drug design based on structural data.	Designing ligands with optimal binding conformations.	
		Protein-ligand interaction analysis.	Predicts binding affinity more accurately.		
Homology modeling: protein and peptide design	The concept that proteins and peptides sharing amino acids sequence similarity typically exhibit similar structural characteristics enabling the design of proteins and peptides with specific interaction or functions.	Sequence alignment.	Cost-effective alternative to experimental structure determination methods like X-ray crystallography.	Designing drugs that target specific proteins.	90,91
		Homology modeling.	Yield highly accurate three-dimensional structures.	Modifying enzymes to perform specific functions.	
		Model refinement.	Rapid design of protein and peptide structures.		
De *novo*-design	Creation of new drug-like compounds that resemble existing molecules to have specific pharmacological properties, leverages the concept that chemical structures shared with known compounds could have similar activities.	Database mining.	Novel drug discovery.	AlvaBuilder is capable of generating novel protein structures with a focus on structural similarity to existing proteins.	92–94
		Pharmacophore model.	Targeted Drug Optimization.	The Schrödinger suite includes various software tools to explore chemical space, design novel compounds, and predict their binding affinities to target proteins.	
		Compound generation.	Reduced cost and time.		

Virtual screening focused on similarity is a computational approach used in drug discovery and chemical research to identify potential hit compounds with chemical structural similarities. This method aims to find compounds within a chemical database likely to have similar biological activities or chemical properties to the reference compound. Figure [2.1_1] shows a general flow diagram of a similarity-based virtual screening methodology.

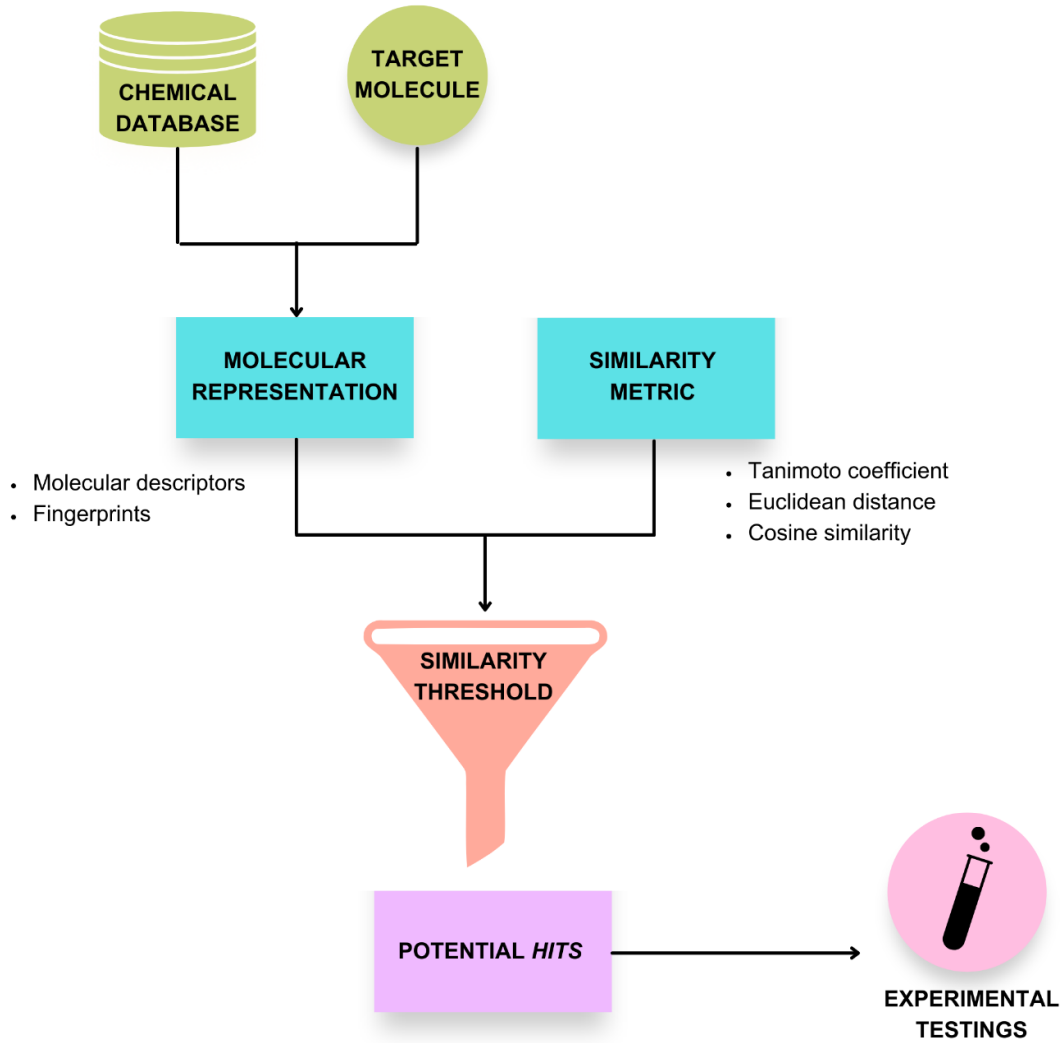


Figure [2.1_1]. Flow diagram of the general methodology for a similarity-based virtual screening approach.

Overall, a similarity-based virtual screening method commences by selecting a target molecule (generally, an established active compound) and the compounds contained in a chemical database. Those elements are commonly represented using molecular descriptors or fingerprints, which capture important chemical and structural characteristics of these molecules. Next, a similarity metric is employed to quantitatively evaluate how similar each compound in the database is to the target molecule. Many similar metrics like the Tanimoto

coefficient, cosine similarity, Euclidean distance, or other relevant measures can be used to assess the similarity based on their molecular representations. Consequently, establishing a similarity value threshold or cutoff value is the further action which sets the minimum acceptable similarity score required for a compound to be considered a potential *hit*. Compounds in the database that surpass the threshold are ranked based on the similarity scores, with those having the highest likeness to the target molecule are given top priority, and are commonly known as potential *hits*. To confirm their biological activity, the compounds identified as potential *hits* through similarity-based virtual screening undergo experimental testing. Experimental validation is a crucial step to verify whether these compounds indeed exhibit the expected biological activity.[95]

Virtual screening can be an iterative process. If the initial set of these *hits* does not have satisfactory results or further refinement is needed, the similarity threshold can be adjusted and rerun the virtual screening or introduce additional criteria to improve *hit* selection. Confirmed *hits* from the experimental validation can be included as new reference compounds in follow-up iterations.

2.2 Extension to other research areas

The underlying principles of molecular and system similarity have been extended to several research areas beyond drug discovery. Some of such extensions are the combination of molecular representations and indices to profile compound data sets, and toxicology prediction with similarity approaches, among others, that will be discussed in the following sections.

2.2.1 Composite indexes

Summarizing the previously described representations, some machine-learning approaches, and indexes have been developed to classify, punctuate, and group molecules according to different categories. Examples are natural product likeness, drug-likeness, and synthetic accessibility or feasibility.

The first approach to compute the likeness of a compound to a natural product set was the Natural Product-likeness Score, developed by Ertl et al. in 2008, which allows for the comparison of the studied compounds with natural products according to their representation of the chemical environment described by molecular fragments, and their frequency of appearance in molecules, compared to natural products and synthetic molecules ref.[96] Some improvements have been made to this approach, to make it more intuitive, portable, and web-based[97]. However, its architecture remains almost unchangeable.

Machine learning has been broadly included in this respect, for example, to classify natural products by their predicted biosynthetic pathway. The first attempt to solve this problem was ClassyFire, which consists of a computational approach to classify molecules according to their taxonomies, in base to > 4800 categories, on a set of unambiguous structural rules, according to the largest structural feature that describes the compound.[98] A most recent example of this kind of application is the NPClassifier neural network, which classifies molecules according to their biosynthetic pathway, superclasses, and classes, recognized by the natural products research community. This goal is achieved by the training of the dataset with labeled information about natural products, described with the counted Morgan fingerprint method for inputting the structural information, which includes the number of atomic substructures, and not only their presence.[99]

Concerning the estimation of a drug-likeness score, many different approaches have been widely used to compute the similarity of molecules to the known compounds used in the clinic, regarding their applicability to virtual screening, associated with the necessity of accomplishing certain parameters or empiric rules, that can dote them with good pharmacokinetic properties.[56,100–102] Those indexes have been developed with different approaches, such as property-based criteria and machine-learning-based approaches. Property-based criteria typically follow the Lipinski and Veber rules, developed more than ten years ago, and there are a variety of different indicators developed with this strategy.[103,104] However, this approach would disregard 16% of currently approved oral drugs, for violating one of those empirical rules.[105] Consequently, these kinds of filters should have been improved with the development of quantitative estimations of drug-likeness (QED), however, evidence suggested that this approach was unsuccessful.[105,106]

Among quantitative approaches to drug-likeness, arises the concept of “molecular beauty” (albeit beauty is, as similarity, a subjective concept), as an estimation of the druggability of a biomolecular target by a chemical compound. The concept tries to integrate the quantitative approach of composite indicators, with the experience and empirical attractiveness of molecules, from the point of view of medicinal chemists.[107] The method consists of ranking compounds on a scale from 0 to 1, weighting the most relevant physicochemical and structural properties, to build the index. Each weighting factor is shown to reflect the importance of that descriptor.

Deep learning approaches have improved the performance of continuous-scaled methods to estimate drug-likeness. Some of them are two class classification training methods[106], and unsupervised learning.[108]

Both of them work by the extraction of structural features, either with graph convolutional networks, or recurrent neural networks, respectively.

2.2.2 Toxicology

Toxicology prediction of chemical compounds in different fields can be initially achieved by the comparison and grouping of chemical structures to those of known toxicological or metabolic reactivity, in function of their physical or chemical properties (recalling the similarity-property principle). Toxicity is a property that can trigger different adverse reactions in the subject and is related to the interaction of the molecules with the organism, e.g. cell receptors, and its relation is mediated by specific interactions between molecules, as well as properties in the chemical entity in the study, categorized in toxicology among the xenobiotic agents. This approach has been implemented in many regulatory agencies worldwide to save resources and time in achieving a first approximation of the chemical safety of a product.[109,110]

Grouping can be achieved by using generic substructural parameters, common functional groups, and probable breakdown products, among other characteristics, and criteria depending on the purpose of the property required to be predicted (i.e., a specific toxical interaction or adverse reaction). Also, the classification can be assessed by a range of doses-response, as the criteria to determine either toxicity or safety.[111] There is a plethora of classic examples of toxicology-related descriptors generalized by similarity approaches, and they need to be harmonized among different legislations to achieve a unified interpretation of similarity criteria and the regulatory validity of results and their interpretation. Discussion on this subject is specifically referred to as the degree of similarity required to consider a compound with probably the same biological properties as those in the evaluated group.[112]

In-silico approaches to assess toxicological information include structure-based and ligand-based methods. Structure-based methods are mainly molecular modeling to identify molecular interactions between compounds and macromolecular targets. Ligand-based methods or *read-across* focus on large databases of active molecules known to produce an effect or interact with a specific target, and can be used to compute physicochemical properties, toxicity, environmental fate, and ecotoxicity.[113] There are diverse software and web-established tools that have been developed based on those rules and criteria, with databases of control well integrated.

2.2.3 Other areas of application

Other areas of common interest have emerged to include the chemical similarity concept among the criteria to preselect chemical compounds to achieve different goals in optimizing biological processes immersed in human activities.

In environmental chemistry, there has been a recent report in which they developed a QSAR model and *read-across* to predict the toxicity of freshwater binary and multicomponent mixtures containing pharmaceutical compounds and pesticides. Their method uses partial least squares to relate toxicity with composed 2D descriptors including atom pairs, fragments, and molecular properties. Through cross-validation, they confirmed the robustness of the model by comparing it with previous reports.[114]

In the dereplication of natural products by using omics techniques, chemical similarity measures have been largely applied. Athayde et al. reported the measure of similarity among ten Brazilian species belonging to the genre of Arnica, according to their metabolomic profiles. The dereplication was carried out using LC-MS/MS, and the representation used for each species was the *m/z* signals of each mixture. Similarity was computed employing different algorithms including squared Euclidian distance, and k-means clustering. Their study reported the natural product composition profile of the Arnica species, and their comparison according to the most significant composites found in them. They also identified previously unreported chemical compounds that belong to species in this category.[115] Skinnider et al. developed an open-source algorithm for the specific computation of similarity among hypothetic natural products. The model is parametrized in the region of the chemical space occupied by natural products. Their method, called Library for the Enumeration of MOdular Natural Structures (LEMONS), employs a representation of circular fingerprints, as well as a retrobiosynthetic approach, that showed a better performance than topological and substructure-based fingerprints. By the use of their method, higher similarity values are assigned to those molecules that can be subproducts of tailored reactions of other natural products, making an efficient assignation of biotransformations among natural products which can be used to taxonomical assignments and research on biosynthetic relationships.[116]

Also, the biosynthetic unexplored chemical space of secondary metabolites from marine prokaryotes has been recently predicted by using molecular similarity according to their biosynthetic gene clusters, diversity,

and novelty, predicting the probable metabolites of such pathways. Tanimoto similarity was computed with a 1D representation of the Morgan fingerprint with a radius of 2. By this approach, they clustered and built taxonomic networks, relating molecules to common substructures, such as scaffolds. Their results showed that 96.8% of the secondary metabolome of marine prokaryotes remains unexplored.[117]

In food science, Sánchez-Ruíz and Colmenarejo reported the potential druggable uses of food chemicals by assessing the similarity coefficients of compounds in FooDB, the largest public database of food chemicals, against compounds in ChEMBL with reported bioactivity against therapeutic targets.[118] Food-drug interactions, such as increasing drug metabolism, decreasing availability, or creating adverse effects, have also been studied by similarity networks. Rahman et al. reported the creation of a machine learning architecture named FDMine, which predicts interactions among food and drug compounds, by computing the closeness of compounds by different approaches of structural ranking and classification by similarity. Their results showed more than 80% of the area under the receiver operating characteristic curve (AUROC). The FDMine approach was tested in the discovery of unknown food and drug interactions.[119]

Besides drug design, there are many other pharmaceutical and clinical applications in which the similarity concept plays a key role. Recently, Ellin et al. reported an algorithm for data mining in biochemical applications of imaging mass spectrometry employing extended similarity indices. Results on hyperspectral data across the studied biological surface (tissue) are processed by principal component analysis (PCA) to compute scoring and loading values for each *m/z* peak. Adopting extended similarity indices, which compare multiple objects simultaneously[120], they proposed a protocol for comparing multiple mass spectra with results of imaging mass spectrometry, with better correlated spectral results, for morphological tissular regions. This approach permits the difference of biological versus non-biological tissue regions, by determining greater values of similarity among them. It also represents biological regions with real lipid m/z peak intensities, that can be rationally interpreted in PCA results.[121]

3. Visualization of similarity

Similarity is a cornerstone concept in molecular design that allows the study of SPRs, as well as the generation and exploration of chemical space. These applications have demonstrated a significant impact in different areas (*vide supra*), and their development has accelerated in recent years.[122] But, similarity is a subjective and “intangible” concept that ideally can be represented with different visualization methods, allowing an

intuitive representation of the concept. For example, Table [3.1_1] summarizes representative strategies that allow their visualization and study resolving different kinds of queries.

Table [3.1_1]. Representative similarity strategies and their applicability domain on molecular design.

General similarity applicability	Specific similarity applicability	Similarity visualization strategy	Query	Ref.
Decoding SPRs	Organization and annotation of substructures	SAR maps	Allows the use of similarity R-groups to identify important chemical motifs to rationalize the studied property.	123
		Scaffold tree	Allows the use of scaffold similarity to identify important chemical motifs to rationalize the studied property.	124
		SAR matrix	Allows the use of molecular pairs similarity to identify important chemical motifs to rationalize the studied property.	125
		Chemical space networks*	Allows the use of molecular networks similarity to identify important chemical motifs to rationalize the studied property.	126
		Constellation plots*	Allows the use of sub-structure similarity to identify analog leads series.	127
	Data mining: Identification of important molecules in a dataset	SAS maps	Allows the identification of “property cliffs” and “scaffold hops” of a specific endpoint.	128
		DAD maps	Allows the identification of “dual property cliffs” and “dual scaffold hops”, and selective compounds against one or two different endpoints.	129
		ChemMaps	Allows the identification of “satellite” molecules that *per se* could represent the totality of a dataset.	130
	Quantification of SPRs	Decodification of supervised learning methods	Allows the identification and interpretation of descriptor features that determine compound activity predictions obtained by machine learning methods.	131
		Quantification of descriptors influence	Allows the quantification of the fingerprint and descriptor's influence to represent SPRs.	132
Construction and decoding chemical space	Dimensional reduction	PCA	Allow reducing the dimensionality and the study of the different representations and descriptors.	14
		SOM		133
		t-SNE		134
Data integration		Shinyheatmap	Allows visualization and analysis of ultra large -omical data.	135
	Consensus similarity	Consensus Diversity Plot	Allows quantifying the consensus similarity of different datasets using different chemical representations and	136

	quantification and visualization		descriptors at the same time.	
* They could also be classified within the category "Data mining: Identification of important molecules in a dataset"				

As shown in Table [3.1_1], the similarity principle could be graphically represented using the “chemical space” concept, defined that "An M-dimensional cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors".[14,137] Namely, is a graphical form to condense and represent high dimensional molecular data, e.g., from molecular representations and descriptors, that are interrelated by their similarity. Multiple reviews exist that remark on their utility and impact on different areas. However, there is not a unique chemical space representation because this depends on their molecular representation (or descriptors) and similarity metric used to construct it. In other words, it is possible to generate numerous representations of the chemical space for the same chemical data set.[138] However, a recent data fusion concept named “consensus chemical space”[139] has shown that it is possible to combine or “fuse” chemical spaces, improving their power chemical description.[140]

Similarity networks are an emerging approach to generate complex data fusions and similarities comparisons[141], useful to reduce the data gaps on large and ultra-large datasets from small (i.e., ligands) to large molecules, i.e., peptides and protein[142,143], and to decode the three-dimensionality features associated with a specific property.[144]

Figure [3.1_1] shows different visualization methods that use the similarity concept to decode properties, substructures, or global similarities between compounds or chemical datasets. For example, panel (A) shows an example of a classical visual representation of the chemical space that facilitates the comparison of two compound data sets (each one represented with blue and red data points). In this example, the compounds are represented by multiple dimensions and t-SNE is used to project the multidimensional space into a 2D plot. The visualization in Figure [3.1_1]B is an example of a consensus diversity plot[136] that enables the comparison of the relative similarity of five data sets based on multiple diversity criteria: fingerprint-based similarity (X-axis), scaffold diversity (Y-axis), property similarity (mapped with a continuous color scale; and dataset size (mapped with the relative size of the data points). Figures [3.1_1]C and D are examples of structure-activity similarity (SAS)[128] ref and dual-activity difference (DAD) maps[129], respectively, which are graphical and quantitative methods to explore the activity landscapes of compound data sets using pairwise

comparisons. SAS and DAD maps enable the rapid identification of activity, property and selectivity cliffs, as well as selectivity switches. Figure [3.1_1]C is an example of a Constellation Plot[127] which is characterized by the visual comparison of the diversity of analog series (represented by each point in the constellation plot), the fingerprint-based similarity of the analog series, plus the amount of compounds populating each series (indicated with the data point size). The continuous color scale facilitates mapping an additional property e.g., biological activity value. Figure [3.1_1]F is an example of a molecular similarity by network to explore the SAR of compound data sets and, like SAS and DAD maps, rapidly identifying activity cliffs.[145]

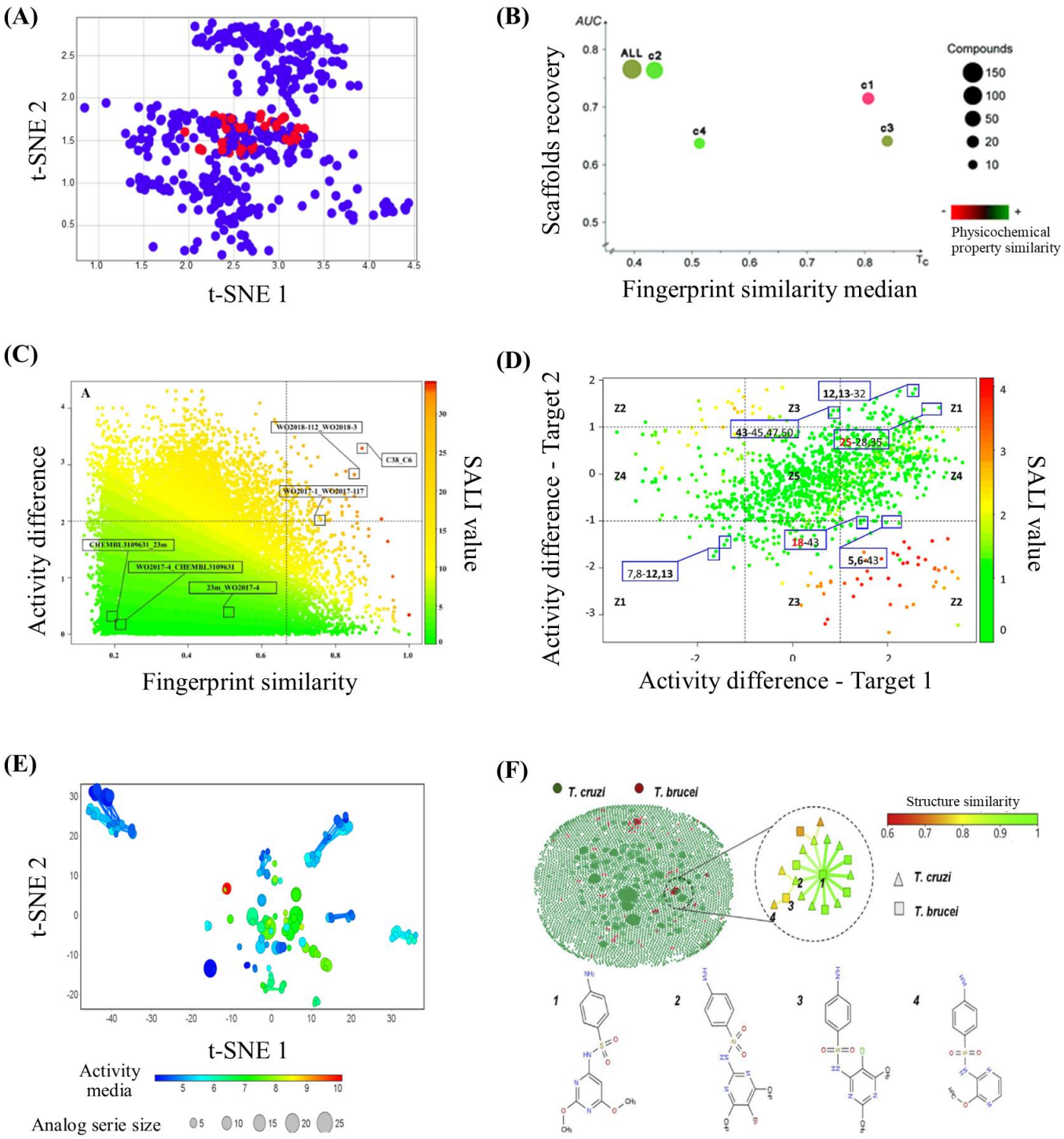
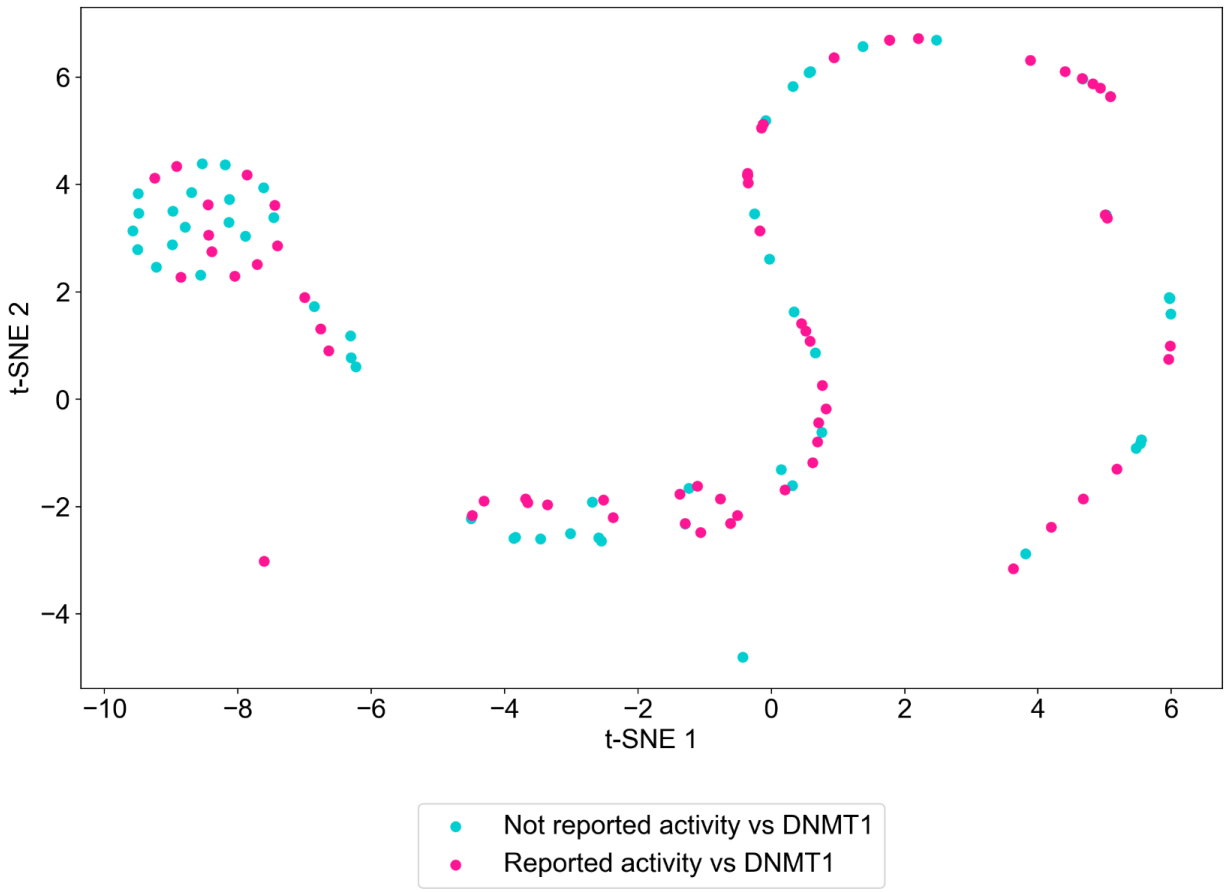


Figure [3.1_1]. Examples of visualization methods that use the similarity principle. (**A**) Dimensional reduction visualization (t-SNE plot); (**B**) Consensus similarity quantification and visualization (Consensus diversity plot); (**C** and **D**) Data mining visualization of paired compounds (SAS and DAD maps); (**E** and **F**) Organization and annotations of substructures visualizations (constellation plot and chemical space networks). Figures adapted from published studies refs.[134,140,145,146]

3.1 Representation of chemical space as works of art

Beyond providing chemical information, visual representations of the chemical space and chemical multiverses can also convey an artistic meaning or expression, that, in turn, might help in the efforts of education in chemoinformatics and scientific dissemination. The authors recently proposed a natural extension of cheminformatics in art through artistic representations of chemical spaces[147] and generated an exemplary Chemical Space Art Gallery publicly available (https://www.difacquim.com/chemical-art-gallery/)[147]. Figure [3.1_1] shows a representative visualization of the chemical space. Figure [3.1_1]A illustrates a typical scientific representation of the chemical space of a database of 120 dietary chemical compounds with reported epigenetic activity in a t-distributed Stochastic Neighbor Embedding (t-SNE) visualization. The dietary compounds are divided in two groups: 1) the compounds with reported activity vs DNA methyltransferase 1 (DNMT1) and 2) not reported activity vs DNMT1. As with many visual representations of the chemical space, the interpretation is driven by the similarity relationships between the chemical properties, in this example, the properties employed for the two dimension reduction analysis were: octanol-water partition coefficient (SlogP), topological polar surface area (TPSA), molecular weight (MW), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA) and rotatable bonds (RB). Figure [3.1_1]B illustrates the “transformation/translation” of the same visualization into a digital painting or artwork. As discussed elsewhere, despite the fact AI is increasingly used in digital artworks, visualizations of chemical spaces are ultimately driven by the concept of molecular similarity.

A



B

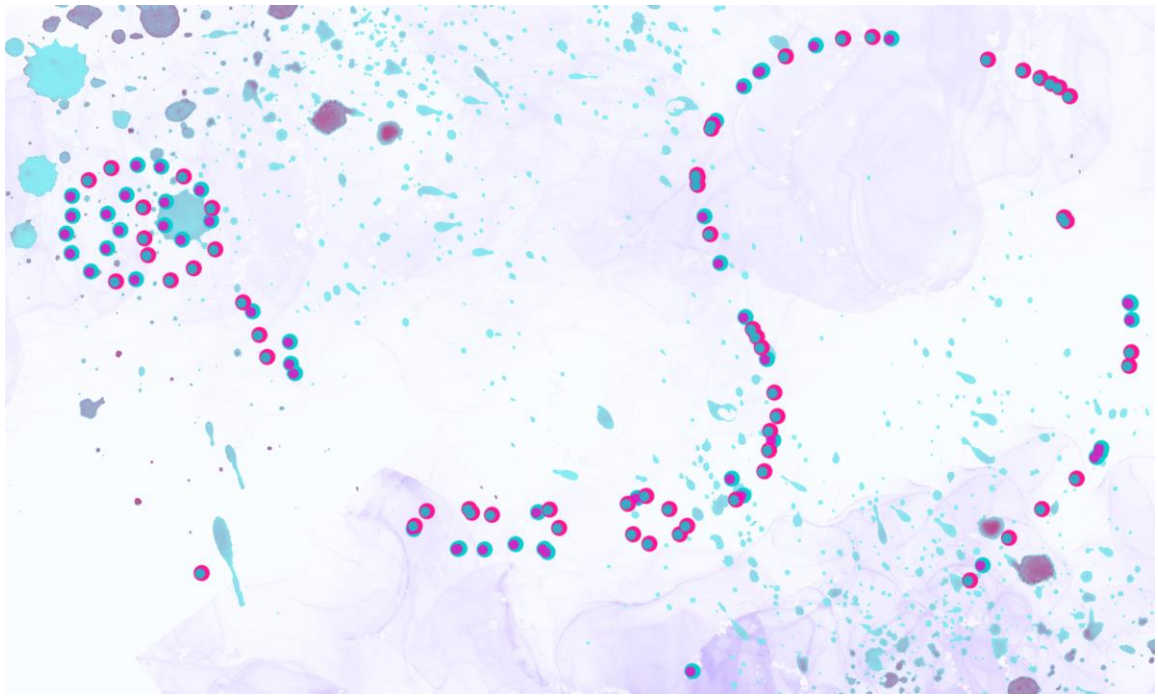


Figure [3.1_1]. **A**) t-SNE visualization of a dietary database comparing epigenetic activity against DNMT1, using drug-like descriptors. **B**) Art work of chemical compounds generated from panel A information.

4. Extended similarity

Until this point of the review we have discussed the similarity measurements in a pairwise manner. While obviously important, this paradigm can be less-than-ideal when we want to analyze very large libraries. Recently, the extended similarity framework was introduced to the field with the goal of alleviating this problem.[120,148] Extended similarity allows the comparison of multiple molecules at the same time, yielding a similarity value for the whole dataset. For a set of *N* molecules represented by binary fingerprints of size *M*, we will have a matrix $N \times M$. The first step to get to the extended similarity value is to sum the matrix column wise to get a vector $\sum = [\sigma_1, \sigma_2, \dots, \sigma_M]$. With the purpose to identify how each column sum, σ_k, contributes to the similarity or dissimilarity of the set, we will define the quantity $\Delta_{\sigma_k} = |2\sigma_k - N|$. We then define a coincidence threshold, γ, that will help to classify each column according to how uniformly distributed the elements of a bit position are. The classification rules are: i) if $2\sigma_k - N > \gamma$ it will be a 1-similarity column, ii) if $N - 2\sigma_k > \gamma$ it will be a 0-similarity column, iii) otherwise it will count as dissimilarity. The final step is to weigh the cases in which we have a partial coincidence (not all the elements are 0 or 1). Conveniently defined weighting functions can be defined to take this non-perfect coincidence into account for partial similarity, f_s, and for partial dissimilarity, f_d. Examples of used weighing functions are:

$$f_s(\Delta_{\sigma_k}) = \frac{\Delta_{\sigma_k}}{N}$$

$$f_d(\Delta_{\sigma_k}) = 1 - \frac{\Delta_{\sigma_k} - N \bmod 2}{N}$$

The mentioned steps can lead to a generalization of pairwise indices. For example, the commonly used Jaccard-Tanimoto index is transformed in the following way:

$$s_{JT} = \frac{a}{a+b+c} \rightarrow s_{JT} = \frac{\sum_{1-s} f_s(\Delta_{\sigma_k})}{\sum_{1-s} 1 + \sum_d 1}$$

Note how the *a* counter transforms in a summation over the 1-similar counters and the dissimilar counters, *b* and *c*, into the respective summation over the dissimilarity columns. This same transformation can be done for any similarity index. The major advantage of extended similarity indices is that it can calculate the overall similarity of a set much more efficiently than by using the traditional pairwise comparisons. With the later methodology, one would need to perform *N(N-1)/2* comparisons, which scales as O(N^2). With the extended similarity framework, the scaling is linear, O(*N*), and as such, dramatically more efficient. This fact opens the door for a plethora of applications for extended similarity measurements. [120,148]

One of the most popular ways of picking a diverse set of molecules from a bigger set is the *MaxMin*[149] algorithm. In this method starting from a random first molecule, the pairwise comparison of the picked compounds and the not picked compounds is computed; the selected compound will be the one with the smallest value for the biggest similarity between itself and the selected compounds. This procedure iterates until the desired number of molecules is picked. The whole method is computationally costly for large number of molecules. A method inspired by the *MaxMin* algorithm but using extended similarity was developed,

named *Max_nDis*. Starting from a random molecule, the next compound to pick will be the one that minimizes the extended similarity when it is added to the set of already picked molecules; it continues until the desired number of molecules is reached. The *Max_nDis* method was proved to be faster and superior method for diversity selection in various datasets; overall the similarity of the picked molecules was always lower than the ones picked with the traditional methos.[148] Extended similarity was used as a performance metric of novel sorting genetic algorithms, NSGA-II and NSGA-III, for small-molecule optimization; an example on how it can be used as a metric for chemical diversity.[150]

The extended similarity framework can also be used to gain information about particular objects in a set. If we calculate the extended similarity on a set excluding one molecule, we get the complementary similarity for that object. The object with the lowest complementary similarity will be the one most similar to the rest of the set, meaning that it will be medoid of the set. Hence, the object with the highest complementary similarity will be the outlier (least similar to the rest) of the set. Ranks based on complementary similarity can help to sample regions of interest of the set. To retain the linear scaling, one can simply subtract the binary vector corresponding to the object, $v_i = [0, 1, \dots, 0]$, from the column wise sum vector $\sum = [\sigma_1, \sigma_2, \dots, \sigma_M]$, resulting in a new vector $v_i - \Sigma$. If we apply the same extended similarity algorithm on this vector we get the complementary similarity value, and retaining O(*N*) scaling even if we do it for all the set.[151]

Another application of extended similarity indexes to diversity is to find relations between large chemical libraries. The extended similarity indices can be used to compare the diversity of two libraries by taking their absolute and relative differences, to then use a sigmoid function to compare if a set is less or more diverse in reference to other.[143] Inspired on Chemical Space Networks (CSNs)[126,152] and with the help of extended similarity, Chemical Library Networks (CLNs) were developed to explore the inter use extended similarity indexes to find the interconnections between libraries' diversities. To generate the CLNs the extended similarity of the union of two sets ($s_e(A \cup B)$) was calculated with a constant coincidence threshold (γ). Then taking each library as node, all the possible edges are generated; then edges are pruned based on if the extended similarity of the union of those two sets is below an edge-threshold. By increasing the edge-threshold one can find the most important diversity connections and find a central library in a set of libraries. The same process can be repeated without edge pruning and changing the coincidence threshold when calculating the extended similarity of the set's union; and using the extended similarity as measurement of the magnitude of the connections.[143] CLNs have been used in a chemoinformatic analysis of libraries focused on epigenic targets. In this study two libraries of 11 were identified as the centers of the network, these two libraries were also the least diverse ones;[153] also for study of DNA-encoded libraries.[154]

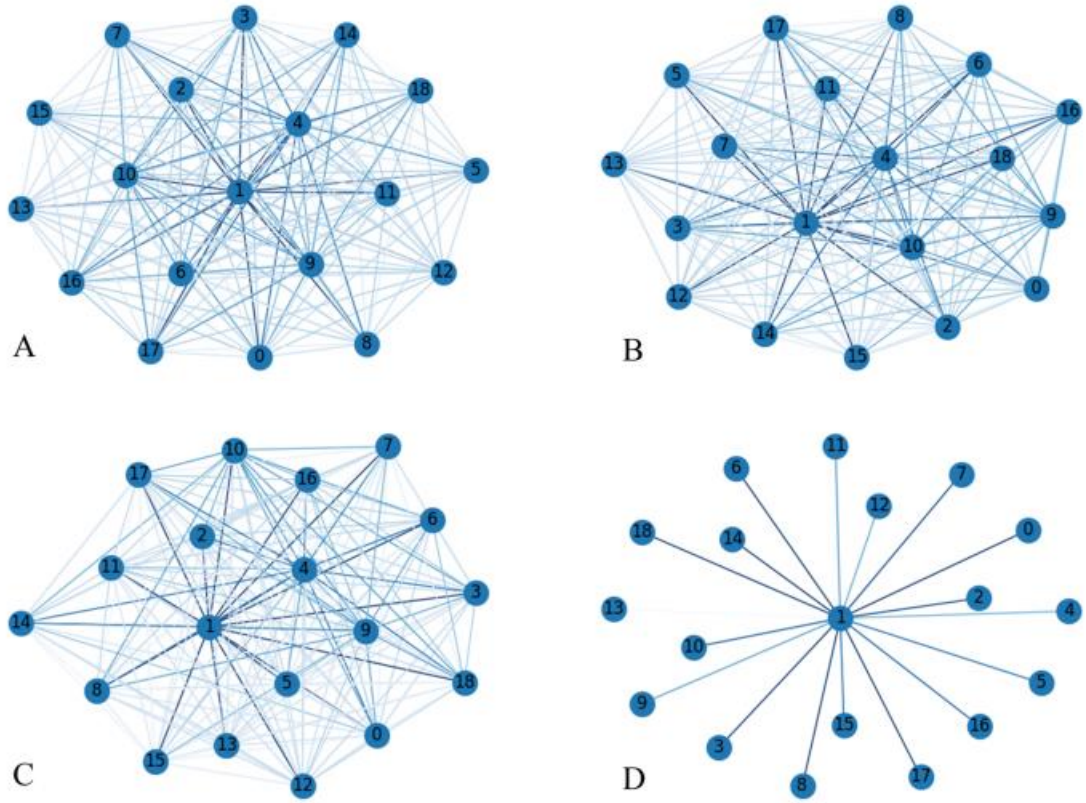


Figure 1: CLNs using RDKit fingerprints with a constant coincidence threshold with (A) all edges and pruning edges at thresholds (B) 0.1, (C) 0.2, and (D) 0.3 for 19 chemical libraries.[143]

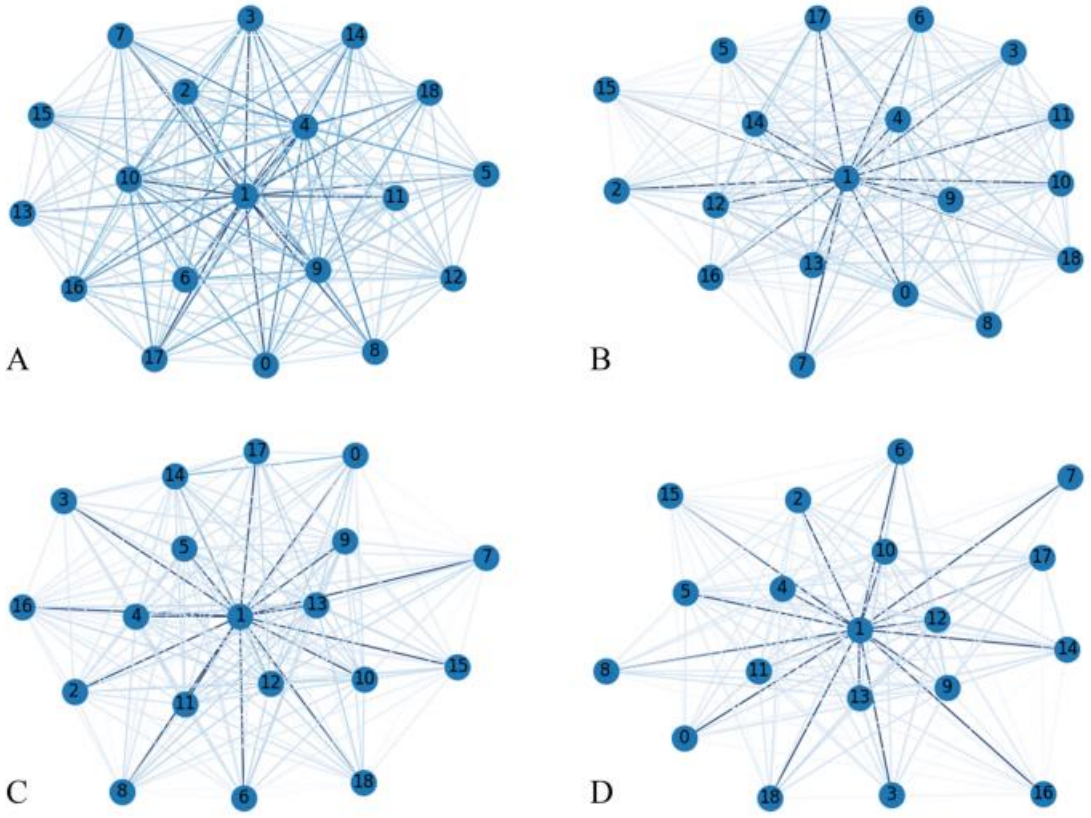


Figure 2: CLNs using RDKit fingerprints with all edges calculated with (A) minimum coincidence threshold, and coincidence thresholds of (B) 30%, (C) 70%, and 90% of the number of elements in the set's union for 19 chemical libraries. [143]

One of the major advances on the extended similarity was advancing beyond binary vectors and generalizing it for categorical[155] and continuous[156] numerical variables. For the latter case, three variants were proposed for the processing of numerical continuous data, for all scaling of the data between 0 and 1 needs to be done a priori. The first variant based on the sum of the absolute difference of values and the mean value of its column $\left|x_{ij} - \bar{x_j}\right|$, the second one is similar to the first one but now doing $1 - \left|x_{ij} - \bar{x_j}\right|$, and the third variant only calculates the sum over the scaled data. In all three cases coincidences measurements are calculated to classify the columns as high-content similarity, low-content similarity or dissimilar.[156] The selection of descriptors is key for Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) models.[157] Extended similarity provides a flexible similarity calculation that can help find the normalization method, coincidence threshold, descriptors and coincidence variant that can separate the similarity of the two groups of compounds the best, for example active and inactive compounds. Applied on a cytochrome P450 2C9 data set, the just explained framework was used select the continuous descriptors that separate the similarity of the inactive and active ligands; by comparison the inactive compounds have a smaller similarity than the active.[156] Extended similarity was also used to evaluate created molecular fingerprints from physicochemical data, this was done as a proof of concept for a set of 13 physiologically active compounds.[158]

Activity cliffs correspond to molecules that are highly similar but have a big difference in a given property. An approach to quantify the activity cliffs in a set is the SALI (Structure-Activity-Landscape-Index) index:

$$SALI(i,j) = \frac{\left|P_i - P_j\right|}{1 - s(i,j)}$$

As it can be appreciated is a pairwise comparison, hence, to quantify the activity landscape of a whole set would require *N(N-1)/2* comparisons. To diminish the computational cost, eSALI was introduced, as given on equation 5. It can be appreciated that the operations required to get the eSALI index scale O(*N*). This new metric serves to quantify the landscape's roughness and modelability of the set with the given representation.[13]

$$eSALI(M) = \frac{1}{N}\frac{\sum|P_i - \bar{P}|}{1 - s_e(M)}$$

Another field that extended similarity can be used as tool is chemical space exploration. With the pairwise similarity matrix of the compounds in a library one can do a PCA to visualize the chemical space, however, is a computationally costly task. ChemMaps is an alternative that uses only a certain portion of the compounds, satellites, that yields a visualization where the distances and form resemble the one with the complete matrix.[130] With complementary similarity calculations it is possible to select the satellites from the medoid and/or periphery region or uniformly. In this way we can identify regions of the library's chemical space that are key to sample to have a good ChemMap. Figure 3 shows one example of a ChemMap.[159]

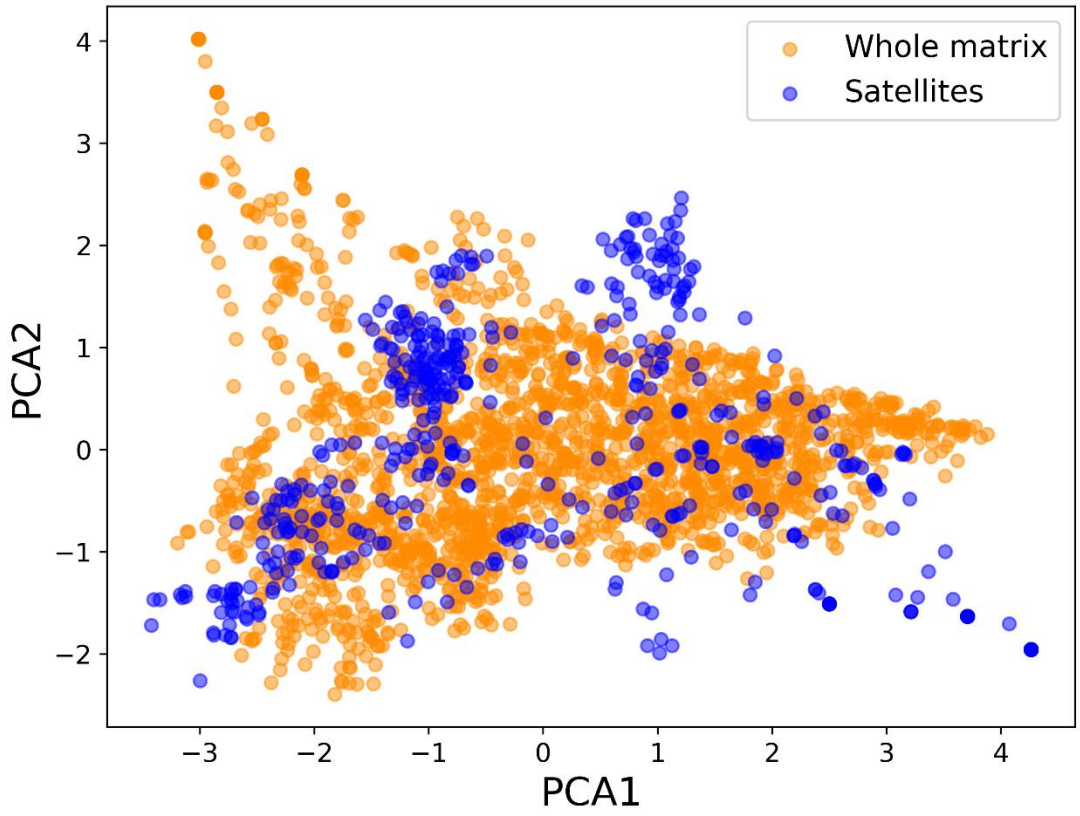


Figure 3: ChemMap of NuBBE [cite] library ECFP4 fingerprint using 25% as satellites sample with medoid-periphery alternations.[159]

Imaging mass spectroscopy has also benefited from extended similarity. Mass spectra can be converted to binary fingerprints, using PCA to identify high, mid and low correlation pixels extended similarity is used to decide if the pixel corresponds to a biological structure or not. With extended similarity medoid spectra can also be extracted to represent groups picked by PCA and help aid the imaging interpretation. It also shows potential applications to the grouping of correlated regions in the tissue by picking important pixel groups from the beginning, complementing PCA calculations.[121]

5. Molecular similarity in Molecular Dynamics simulations

The importance of molecular comparisons, and the ever-present need to develop more efficient algorithms to perform them is particularly evident in the field of Molecular Dynamics (MD), a computer simulation method for studying systems' dynamics by integrating Newton's 2nd Law. Although the advent of graphical processing units has made microsecond timescale MD simulations a routine, simulation post-processing analysis failed to keep pace with the increasing size of simulation datasets.[160] Here, comparisons are important mainly for their role in unsupervised learning techniques. This has been crucial in this area, enabling dimensionality reduction for faster analysis and the grouping of similar samples through clustering. Unlike supervised learning, unsupervised learning relies solely on input information to extract patterns and structures, making it

useful for uncovering insights in complex datasets. Unsupervised learning has a versatile of applications to discover patterns, anomalies, and relationship within data. In Molecular Dynamics, defining collective variables (CV), which are a set of pre-defined features that guides enhanced sampling, is a challenge because they are highly sensitive to user choices and require dimension reduction techniques for accurate representation.[161] CV that represents the metastable states of the systems' dynamics is more likely to find the path of the slow modes in protein or protein-ligand dynamics.

In the following section, we will talk about how to select features for representing the protein or protein ligand systems, dimension reduction techniques for representing high-dimensional data, different clustering methods, and lastly, examine the role of extended similarity in Molecular Dynamics analysis.

5.1 Featurization

Featurization is the selection of the most represented features of a data. The most intuitive features would be the spatial coordinates of atoms in the trajectory. However, this can be challenging because configurations need to be aligned to some reference(s) to minimize the effect of local rotations and translations during a simulation and the decision of what reference(s) to align to can impact your data analysis significantly.[162,163] To circumvent this limitation, distances, dihedrals, or angles of atoms in every frame, is another way to represent biomolecules.[164] Alternatively, there have been many methods of representing protein-ligand interactions in fingerprints. Structural based protein-ligand interaction fingerprints encode interaction in binary vector. Each residue is encoded in a vector, which consists of information such as hydrogen bonds, hydrophobic interactions, aromatic stacking, ionic bond, the distance/angle of the interaction, and the atoms and residues involved.[49,165] PROLIF is software for generating fingerprints for representing molecular interaction between combination of protein, ligand, RNA, and DNA molecules, which is compatible with MD trajectories, docking, and experimental structures.[166] Energy-based interaction fingerprints can be useful to understand the intermolecular interaction during the dissociation mechanism when a ligand is dissociated from a protein.[167] The interaction fingerprint would be useful if you need cluster binding different modes of the protein-ligand complex, starting ensemble for docking, and dissociation mechanism studies.

5.2 Dimension reduction techniques

With an increasing surge of data, strategies to project high-dimension data into low-dimension data for data analysis are needed to process the data using the least amount of information. Dimension reduction (DR) is critical in clustering, or the classification of similar data into its own group. It can not only reduce the computational cost of clustering but can also trim the noise of data, which can improve the accuracy of clustering results. DR is not limited to applications of clustering; it also has applications in Molecular Dynamics in areas such as enhanced sampling and building Markov State Models (MSM). For applications to enhanced sampling, a good low-dimension projection of the data is needed to characterize collective variables (CV), which can better guide the enhanced sampling to bias potential to the Hamiltonian of the system or be used to build an MSM.[168] Dimension reduction is divided into linear and nonlinear.

Linear dimension reduction methods are helpful to define collective variables as they create a linear combination of the most important principal components. Principal component analysis (PCA)[169] is the most widely used dimensional reduction technique that reduces to 2-dimensional space by projecting the multi-dimension data while preserving the variation of the data. The eigenvector with the highest eigenvalue will be your first PC. It will identify the principal components, which are orthogonal linear combination of the variables. Similar to PCA, time-independent component analysis (tICA)[170] projections involves a linear transformation of the input data, aiming to maximize the autocorrelations in the output data. Because of this, it is widely used in applications to MD because it identifies the slow motions in an MD simulation, while preserving kinetic information.[168,171] Both PCA and tICA has been to build MSM and in a study conducted to compare the performance of both techniques, both were able to enhance the performance of MSM.[171] However, it is noted that both methods suffer significantly with an increase in dimensions. Nonlinear DR technique includes t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) map distances or similarities between data point in high-dimensional space to a low-dimensional space. Therefore, it preserves the relationships and structures based on how close or similar points are.

In one study comparing 2D and transcriptomics datasets, t-SNE and UMAP are better at preserving local structures (preservation of neighbors), while PCA is better at preserving global structure (relative position of different clusters) and less sensitive to parameter choices.[172] A similar sentiment is shared in a MD simulation, in which PCA and tICA perform the best to capture the slow modes of protein dynamics and the free energy surface of small peptide and protein.[173–175] While t-SNE preserves similarity between high- and low-dimensional space, at the same time, it performs poorly in extraction of global structures. Alternatively, it was found that UMAP overall not only performs better in retaining kinetic information and Cartesian coordinates than t-SNE, PCA, and tICA[168], but also performs better in MSM and able to resolve more metastable states.[176] All methods have their advantages and limitations, and the choice of how many dimensions is needed, choice of parameters, preprocessing steps, and DR technique choice is system dependent and requires careful decisions.

5.3 Clustering

Clustering is the grouping of similar samples together, which is essential for unraveling protein folding dynamics, constructing Markov State Models, enhancing Replica Exchange simulations, and discerning drug binding modes. Root-mean square deviation (RMSD) is a widely used pairwise metric for calculating similarity between configurations in Molecular Dynamics simulations and popular among clustering metrics. Below we will investigate several widely used algorithms, which differ vastly in their approaches, time complexity, sensitivity to outliers, parameters, and effectiveness. It is dependent on the preprocessing decisions, parameter chosen, metric for comparison, and algorithm chosen.

K-means is known to be the most popular partitioning clustering algorithm and its input requires two parameters, the number of centroids, and initial estimate of centroids. From these starting initial centroid estimates, it assigns each point to the closest centroid, and recomputing the centroid of the newest formed clustering. This process is iterated until convergence, when the computed centroid remains the same and the

points remain in the same clusters. When applied to Molecular Dynamics, it proves to be efficient as it has a time complexity of O(k * N * i), k for number of cluster, n for number of point, and i for number of iteration. However, it fails to identify key metastable states and fails short for non-convex cluster shapes.[177] In addition, k-means clustering is dependent on the initial centroid estimation. In the standard method of initiating centroid, k-means++, it picks a random point and picks the maximally distant points for the k number of clusters. Another major pitfall is the decision how many clusters to pick. This is not a trivial task due to the complex nature of MD simulation. Users would require prior knowledge about the system to make a reasonable conjecture. Due to its partitioning nature of assigning points to its closest centroid, its cluster shapes are mainly uniform and highly sensitive to outliers. K-means is more useful for data that is well-separated. K-medoid follows the same algorithm as K-mean with the main difference being that pairwise comparison has to been to calculate the similarity between every point to determine the least dissimilar object, marking the time complexity to be O(k * N^2 * i).

Hierarchical agglomerative clustering (HAC) creates a cluster for every individual and begins merging similar clusters together. There are several common merging techniques. Single linkage merges clusters based on the distance between the two closest points of two clusters. Complete linkage merge clusters based on the maximum distance between two clusters. Centroid linkage merges clusters based on the distance between the two closest centroids of two clusters. Average linkage merges clusters based on the average distance between all combination of points of two clusters. Average linkage and centroid perform the best as it not only produces compact and different clusters, but also identify different shapes and sizes; this come with the pitfall of being most time intensive.[178] Although computationally efficient, single linkage tends to add long tails of data to one cluster[179] and complete linkage tends to have smaller clusters. Some advantages of HAC include: no prior knowledge needed on number of clusters; parameter independent. Some limitations of hierarchical include: a time complexity of O(N^2) due to the pairwise similarity computed at every iteration[173,180], sensitivity to outlier, and variant to permutation in the data.[178]

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[181] alleviates some of the pitfalls in K-means, which is the inability to partition noisy data, limited to globular clusters and similar shaped clusters, and the need to specify number of clusters. DBSCAN is a density-based clustering method that can identify highly dense regions and clusters of arbitrary shapes and can identify metastable state.[177] DBSCAN takes in two parameters to define density, radius for forming epsilon neighborhoods and minimum number of samples. DBSCAN is a density-based algorithm. First every point will form their epsilon neighborhoods for all the samples within a radial threshold. Epsilon neighborhood that meets the minimum number of samples will be considered a core sample of high density. Then a random core point is chosen to begin the first cluster. Samples in its epsilon neighborhoods will be included in the first cluster and the samples on the cluster periphery will look for overlapping epsilon neighborhoods until there are no more epsilon neighborhoods. Lastly, points close by to the first clusters will also be included in the cluster. This process will continue for other clusters. This process will be able to distinguish high- and low-density regions. How does it affect MD trajectories? Using only these two criteria for clustering, it can lead to small clusters and be susceptible to noises. There DBSCAN requires a careful selection of parameters as it is very sensitive to epsilon and the

minimum number of points. In addition, density-based clustering suffers from a large performance cost with the best time complexity to be O(nlogn)[182] and the worst time complexity of $O(N^2)$, every point needs to be compared to every other point to find its neighbors.[183] This can come as a cost when data is high dimensional. Although DBSCAN is susceptible to noise. It can fail in a dataset with varying densities, or the data is highly complex. Therefore, HDBSCAN[181] is an extension to DBSCAN as hierarchical method that can combine clusters to determine the optimal number of clusters based on the stability of the clusters. DBSCAN requires a static threshold as the input. However, HDBSCAN uses a range of different thresholds at every iteration and can form or merge clusters depending on the hierarchical structure.[184] This improves DBSCAN by the dissolution of the epsilon parameter. The pitfall of DBSCAN lies in its sensitivity to parameters, which requires tests. HDBSCAN overcame this limitation by involving methods that try out all possible thresholds and produce high-density clusters.

5.4 Extended Similarity in MD simulations

Extended similarity indices were able to distinguish conformations for a protein folding mechanism. It successfully identified two distinct protein folding pathways with hierarchical agglomerative methods.[151] However, because extended similarity indices were computed for a binary vector input in the form of a contact map, this is effectively scaling $O(N^2)$, which means more computation in the preprocessing step to featurize the data to be compatible with extended similarity indices. The feature selection, as discussed previously, is highly dependent on the system and requires prior knowledge.

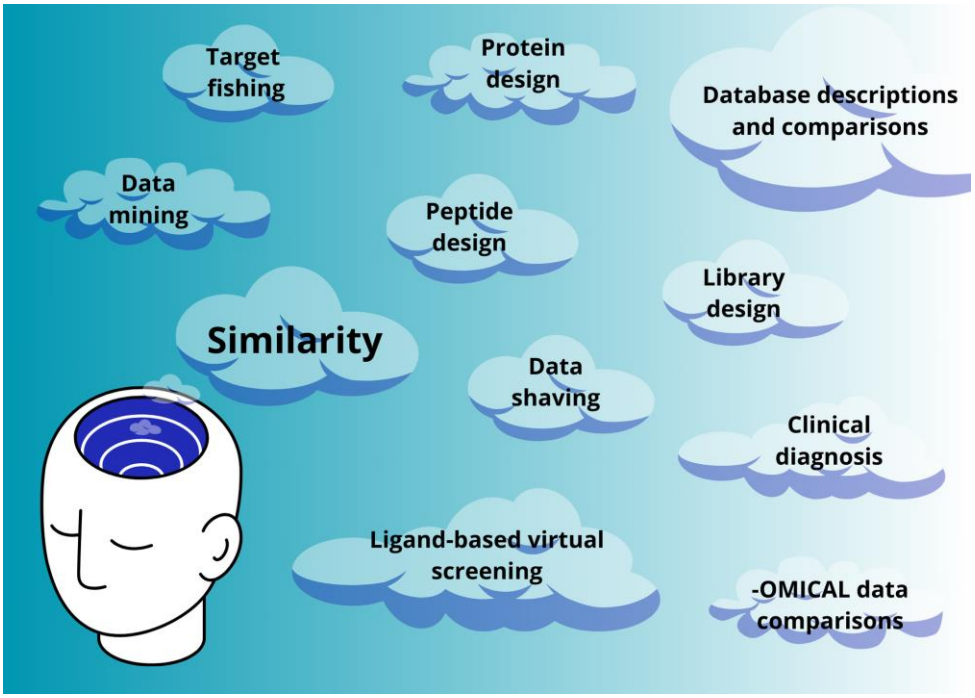
Extended continuous similarity is an extension to extended similarity. It takes the Cartesian coordinates, matrix of number of samples or frames times number of features, as the input for comparisons.[156,160] The samples will be condensed into a single vector containing the column-wise sum of features for all the samples. For matrix of Cartesian coordinates to be compatible with extended similarity indices, it needs to be normalized between [0,1] to accommodate the range of most cheminformatics metrics. To do this, we normalize between the minimum and maximum of the coordinates, which is to be compatible with the ranking of RMSD calculations.

Build efficient and general clustering algorithms. Root-mean-square-deviation (RMSD) is the main tool for calculating the similarity of MD frames due to its straightforward implementation. However, it has limitations as a gauge in cluster conformation; it compares in pairs, effectively making it an $O(N^2)$ algorithm; and this similarity metric is influenced by other parts, especially ligands or flexible regions. fraction of candidates from high-density regions in the data, then use our diversity algorithms (Max_nDis, ECS_MeDiv) to select representative centroids at a linear scale.[148] Diversity selection builds upon ideas from extended similarity. First it selects a seed or the starting point for diversity selection. Then it iterates through the data and picks the next point that gives the maximum extended similarity. This method not only provides a method of identifying a diverse set of distinct conformations, but it is also deterministic and linear scaling.[160] From this study, frames in a MD simulation are more diverse when selected with the diversity algorithms than using RMSD.

Acknowledgements

RAMQ, KLP, and LC thank the National Institute of General Medical Sciences of the National Institutes of Health for support under award number R35GM150620.

TABLE OF CONTENTS



References

(1) Miranda-Quintana, R. A.; Kim, T. D.; Heidar-Zadeh, F.; Ayers, P. W. On the Impossibility of Unambiguously Selecting the Best Model for Fitting Data. *J Math Chem* **2019**, *57* (7), 1755–1769. https://doi.org/10.1007/s10910-019-01035-y.

(2) Miranda-Quintana, R.-A.; Cruz-Rodes, R.; Codorniu-Hernandez, E.; Batista-Leyva, A. J. Formal Theory of the Comparative Relations: Its Application to the Study of Quantum Similarity and Dissimilarity Measures and Indices. *J Math Chem* **2010**, *47* (4), 1344–1365. https://doi.org/10.1007/s10910-009-9658-6.

(3) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Differential Consistency Analysis: Which Similarity Measures Can Be Applied in Drug Discovery? *Mol Inform* **2021**, *40* (7). https://doi.org/10.1002/minf.202060017.

(4) Holyoak, K. J.; Morrison, R. G. *The Cambridge Handbook of Thinking and Reasoning*; Cambridge University Press, 2005.

(5) Heit, E.; Rubinstein, J. Similarity and Property Effects in Inductive Reasoning. *J Exp Psychol Learn Mem Cogn* **1994**, *20* (2), 411–422. https://doi.org/10.1037/0278-7393.20.2.411.

(6) Sheridan, R.; Kearsley, S. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discov Today* **2002**, *7* (17), 903–911. https://doi.org/10.1016/S1359-6446(02)02411-X.

(7) Crosland, M. P. The Use of Diagrams as Chemical 'Equations' in the Lecture Notes of William Cullen and Joseph Black. *Ann Sci* **1959**, *15* (2), 75–90. https://doi.org/10.1080/00033795900200088.

(8) American Chemical Society; Academie des Sciences de l'Institut de France; Societe Francaise de Chimie. *The Chemical Revolution*; Paris, 1999.

(9) Döbereiner, J. W. Versuch Zu Einer Gruppirung Der Elementaren Stoffe Nach Ihrer Analogie. *Ann Phys* **1829**, *91* (2), 301–307. https://doi.org/10.1002/andp.18290910217.

(10) Mendeleev, D. On the Relationship of the Properties of the Elements to Their Atomic Weights. *Zeitschrift für Chemie* **1869**, *12*, 405–406.

(11) Wheeler, Q. D. Taxonomic Triage and the Poverty of Phylogeny. *Philos Trans R Soc Lond B Biol Sci* **2004**, *359* (1444), 571–583. https://doi.org/10.1098/rstb.2003.1452.

(12) Engström, E.; Petersen, K.; Ali, N. bin; Bjarnason, E. SERP-Test: A Taxonomy for Supporting Industry–Academia Communication. *Software Quality Journal* **2017**, *25* (4), 1269–1305. https://doi.org/10.1007/s11219-016-9322-x.

(13) Dunn, T. B.; López-López, E.; Kim, T. D.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Exploring Activity Landscapes with Extended Similarity: Is Tanimoto Enough? *Mol Inform* **2023**, *42* (7). https://doi.org/10.1002/minf.202300056.

(14) Medina-Franco, J. L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B. I. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J Comput Aided Mol Des* **2022**, *36* (5), 341–354. https://doi.org/10.1007/s10822-021-00399-1.

(15) López-López, E.; Robles, O.; Plisson, F.; Medina-Franco, J. L. Mapping the Structure–Activity Landscape of Non-Canonical Peptides with MAP4 Fingerprinting. *Digital Discovery* **2023**, *2* (5), 1494–1505. https://doi.org/10.1039/D3DD00098B.

(16) Plisson, F.; Ramírez-Sánchez, O.; Martínez-Hernández, C. Machine Learning-Guided Discovery and Design of Non-Hemolytic Peptides. *Sci Rep* **2020**, *10* (1), 16581. https://doi.org/10.1038/s41598-020-73644-6.

(17) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J Med Chem* **2014**, *57* (8), 3186–3204. https://doi.org/10.1021/jm401411z.

(18) Hönig, S. M. N.; Lemmen, C.; Rarey, M. Small Molecule Superposition: A Comprehensive Overview on Pose Scoring of the Latest Methods. *WIREs Computational Molecular Science* **2023**, *13* (2). https://doi.org/10.1002/wcms.1640.

(19) López-López, E.; Rabal, O.; Oyarzabal, J.; Medina-Franco, J. L. Towards the Understanding of the Activity of G9a Inhibitors: An Activity Landscape and Molecular Modeling Approach. *J Comput Aided Mol Des* **2020**, *34* (6), 659–669. https://doi.org/10.1007/s10822-020-00298-x.

(20) Lewis, G. N. The Atom and the Molecule. *J Am Chem Soc* **1916**, *38* (4), 762–785. https://doi.org/10.1021/ja02261a002.

(21) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J Cheminform* **2020**, *12* (1), 56. https://doi.org/10.1186/s13321-020-00460-5.

(22) Brammer, J. C.; Blanke, G.; Kellner, C.; Hoffmann, A.; Herres-Pawlis, S.; Schatzschneider, U. TUCAN: A Molecular Identifier and Descriptor Applicable to the Whole Periodic Table from Hydrogen to Oganesson. *J Cheminform* **2022**, *14* (1), 66. https://doi.org/10.1186/s13321-022-00640-5.

(23) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J Med Chem* **2020**, *63* (16), 8705–8722. https://doi.org/10.1021/acs.jmedchem.0c00385.

(24) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans Pattern Anal Mach Intell* **2013**, *35* (8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

(25) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(26) RDKit: Open-Source Cheminformatics. Https://Www.Rdkit.Org.

(27) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **2015**, *7* (1), 23. https://doi.org/10.1186/s13321-015-0068-4.

(28) Warr, W. A. Many InChIs and Quite Some Feat. *J Comput Aided Mol Des* **2015**, *29* (8), 681–694. https://doi.org/10.1007/s10822-015-9854-3.

(29) Daylight Chemical Information Systems, Inc. *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

(30) Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and Applications of Chemical Fingerprint for Hit and Lead Screening. *Drug Discov Today* **2022**, *27* (11), 103356. https://doi.org/10.1016/j.drudis.2022.103356.

(31) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J Chem Inf Comput Sci* **2003**, *43* (1), 144–154. https://doi.org/10.1021/ci025516b.

(32) Chu, K. C.; Feldmann, R. J.; Shapiro, M. B.; Hazard, G. F.; Geran, R. I. Pattern Recognition and Structure-Activity Relation Studies. Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System. *J Med Chem* **1975**, *18* (6), 539–545. https://doi.org/10.1021/jm00240a001.

(33) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J Chem Inf Model* **2007**, *47* (6), 2098–2109. https://doi.org/10.1021/ci700200n.

(34) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res* **2017**, *45* (D1), D955–D963. https://doi.org/10.1093/nar/gkw1118.

(35) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(36) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-Fingerprints Detect Similar Activity of Receptor Ligands Previously Recognized Only by Three-Dimensional Pharmacophore-Based Methods. *J Chem Inf Comput Sci* **2001**, *41* (2), 394–401. https://doi.org/10.1021/ci000305x.

(37) Downs, G. M.; Barnard, J. M. Techniques for Generating Descriptive Fingerprints in Combinatorial Libraries. *J Chem Inf Comput Sci* **1997**, *37* (1), 59–61. https://doi.org/10.1021/ci960091c.

(38) Schwartz, J.; Awale, M.; Reymond, J.-L. SMIfp (SMILES Fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules. *J Chem Inf Model* **2013**, *53* (8), 1979–1989. https://doi.org/10.1021/ci400206h.

(39) Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **2006**, *9 3*, 199–204.

(40) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J Chem Inf Model* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(41) Hutchinson, S. T.; Kobayashi, R. Solvent-Specific Featurization for Predicting Free Energies of Solvation through Machine Learning. *J Chem Inf Model* **2019**, *59* (4), 1338–1346. https://doi.org/10.1021/acs.jcim.8b00901.

(42) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J Chem Inf Comput Sci* **2004**, *44* (5), 1708–1718. https://doi.org/10.1021/ci0498719.

(43) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J Med Chem* **2004**, *47* (26), 6569–6583. https://doi.org/10.1021/jm049611i.

(44) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr Med Chem* **2001**, *8* (13), 1573–1588. https://doi.org/10.2174/0929867013371923.

(45) Awale, M.; Reymond, J.-L. Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of ZINC and GDB-17. *J Chem Inf Model* **2014**, *54* (7), 1892–1907. https://doi.org/10.1021/ci500232g.

(46) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J Chem Inf Comput Sci* **1987**, *27* (2), 82–85. https://doi.org/10.1021/ci00054a008.

(47) Daylight Chemical Information Systems. *Fingerprints - Screening and Similarity*. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

(48) Méndez-Lucio, O.; Kooistra, A. J.; Graaf, C. de; Bender, A.; Medina-Franco, J. L. Analyzing Multitarget Activity Landscapes Using Protein–Ligand Interaction Fingerprints: Interaction Cliffs. *J Chem Inf Model* **2015**, *55* (2), 251–262. https://doi.org/10.1021/ci500721x.

(49) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J Med Chem* **2004**, *47* (2), 337–344. https://doi.org/10.1021/jm030331x.

(50) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35* (8), 1334–1341. https://doi.org/10.1093/bioinformatics/bty757.

(51) Vainio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular Overlay Based on Shape and Electrostatic Potential. *J Chem Inf Model* **2009**, *49* (2), 492–502. https://doi.org/10.1021/ci800315d.

(52) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley, 2009. https://doi.org/10.1002/9783527628766.

(53) Grisoni, F.; Consonni, V.; Todeschini, R. Impact of Molecular Descriptors on Computational Models; 2018; pp 171–209. https://doi.org/10.1007/978-1-4939-8639-2_5.

(54) Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure–Activity Applications: A Hands-On Approach; 2018; pp 3–53. https://doi.org/10.1007/978-1-4939-7899-1_1.

(55) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv Drug Deliv Rev* **1997**, *23* (1–3), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(56) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J Med Chem* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.

(57) CAS. *CAS Reactions*. https://www.cas.org/cas-data/cas-reactions.

(58) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol Inform* **2014**, *33* (6–7), 469–476. https://doi.org/10.1002/minf.201400052.

(59) Saldívar-González, F. I.; Navarrete-Vázquez, G.; Medina-Franco, J. L. Design of a Multi-Target Focused Library for Antidiabetic Targets Using a Comprehensive Set of Chemical Transformation Rules. *Front Pharmacol* **2023**, *14*. https://doi.org/10.3389/fphar.2023.1276444.

(60) Daylight Summer School. *Reaction SMILES and SMIRKS*.

(61) Daylight Chemical Information Systems, Inc. *SMIRKS - A Reaction Transform Language*. https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html.

(62) Grethe, G.; Goodman, J. M.; Allen, C. H. International Chemical Identifier for Reactions (RInChI). *J Cheminform* **2013**, *5* (1), 45. https://doi.org/10.1186/1758-2946-5-45.

(63) de Luca, A.; Horvath, D.; Marcou, G.; Solov'ev, V.; Varnek, A. Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches. *J Chem Inf Model* **2012**, *52* (9), 2325–2338. https://doi.org/10.1021/ci300149n.

(64) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. In *Computers in Chemistry*; Springer-Verlag: Berlin/Heidelberg; pp 19–64. https://doi.org/10.1007/BFb0051317.

(65) Maiti, S.; Ram, S.; Pal, S. Extension of Ugi's Scheme for Model-Driven Classification of Chemical Reactions. *International Journal of Chemoinformatics and Chemical Engineering* **2015**, *4* (1), 26–51. https://doi.org/10.4018/IJCCE.2015010103.

(66) Siani, M. A.; Weininger, D.; Blaney, J. M. CHUCKLES: A Method for Representing and Searching Peptide and Peptoid Sequences on Both Monomer and Atomic Levels. *J Chem Inf Comput Sci* **1994**, *34* (3), 588–593. https://doi.org/10.1021/ci00019a017.

(67) Siani, M. A.; Weininger, D.; James, C. A.; Blaney, J. M. CHORTLES: A Method for Representing Oligomeric and Template-Based Mixtures. *J Chem Inf Comput Sci* **1995**, *35* (6), 1026–1033. https://doi.org/10.1021/ci00028a012.

(68) Chen, W. L.; Leland, B. A.; Durant, J. L.; Grier, D. L.; Christie, B. D.; Nourse, J. G.; Taylor, K. T. Self-Contained Sequence Representation: Bridging the Gap between Bioinformatics and Cheminformatics. *J Chem Inf Model* **2011**, *51* (9), 2186–2208. https://doi.org/10.1021/ci2001988.

(69) Zhang, T.; Li, H.; Xi, H.; Stanton, R. V.; Rotstein, S. H. HELM: A Hierarchical Notation Language for Complex Biomolecule Structure Representation. *J Chem Inf Model* **2012**, *52* (10), 2796–2806. https://doi.org/10.1021/ci3001925.

(70) Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. WURCS: The Web3 Unique Representation of Carbohydrate Structures. *J Chem Inf Model* **2014**, *54* (6), 1558–1566. https://doi.org/10.1021/ci400571e.

(71) Matsubara, M.; Aoki-Kinoshita, K. F.; Aoki, N. P.; Yamada, I.; Narimatsu, H. WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. *J Chem Inf Model* **2017**, *57* (4), 632–637. https://doi.org/10.1021/acs.jcim.6b00650.

(72) Johnson, M. A.; Maggiora, G. M.; others. Concepts and Applications of Molecular Similarity. *(No Title)* **1990**.

(73) Maggiora, G.; Medina-Franco, J. L.; Iqbal, J.; Vogt, M.; Bajorath, J. From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *J Chem Inf Model* **2020**, *60* (12), 5873–5880. https://doi.org/10.1021/acs.jcim.0c01249.

(74) Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J Chem Inf Model* **2006**, *46* (4), 1535–1535. https://doi.org/10.1021/ci060117s.

(75) Medina-Franco, J. L. Activity Cliffs: Facts or Artifacts? *Chem Biol Drug Des* **2013**, *81* (5), 553–556. https://doi.org/10.1111/cbdd.12115.

(76) Stumpfe, D.; Hu, H.; Bajorath, J. Advances in Exploring Activity Cliffs. *J Comput Aided Mol Des* **2020**, *34* (9), 929–942. https://doi.org/10.1007/s10822-020-00315-z.

(77) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4* (11), 14360–14368. https://doi.org/10.1021/acsomega.9b02221.

(78) Kumar, A.; Zhang, K. Y. J. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Front Chem* **2018**, *6*. https://doi.org/10.3389/fchem.2018.00315.

(79) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat Rev Drug Discov* **2004**, *3* (11), 935–949. https://doi.org/10.1038/nrd1549.

(80) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J Med Chem* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.

(81) Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; Norris, A.; Sanseau, P.; Cavalla, D.; Pirmohamed, M. Drug Repurposing: Progress, Challenges and Recommendations. *Nat Rev Drug Discov* **2019**, *18* (1), 41–58. https://doi.org/10.1038/nrd.2018.168.

(82) Li, J.; Zheng, S.; Chen, B.; Butte, A. J.; Swamidass, S. J.; Lu, Z. A Survey of Current Trends in Computational Drug Repositioning. *Brief Bioinform* **2016**, *17* (1), 2–12. https://doi.org/10.1093/bib/bbv020.

(83) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform* **2010**, *29* (6–7), 476–488. https://doi.org/10.1002/minf.201000061.

(84) Sun, H.; Tawa, G.; Wallqvist, A. Classification of Scaffold-Hopping Approaches. *Drug Discov Today* **2012**, *17* (7–8), 310–324. https://doi.org/10.1016/j.drudis.2011.10.024.

(85) Schneider, P.; Schneider, G. De Novo Design at the Edge of Chaos. *J Med Chem* **2016**, *59* (9), 4077–4086. https://doi.org/10.1021/acs.jmedchem.5b01849.

(86) Sánchez-Cruz, N.; Fernandez-de Gortari, E.; Medina-Franco, J. L. Editorial: Computational Chemogenomics: In Silico Tools in Pharmacological Research and Drug Discovery. *Front Pharmacol* **2023**, *14*. https://doi.org/10.3389/fphar.2023.1150869.

(87) Liu, M.; Bienfait, B.; Sacher, O.; Gasteiger, J.; Siezen, R. J.; Nauta, A.; Geurts, J. M. W. Combining Chemoinformatics with Bioinformatics: In Silico Prediction of Bacterial Flavor-Forming Pathways by a Chemical Systems Biology Approach "Reverse Pathway Engineering." *PLoS One* **2014**, *9* (1), e84769. https://doi.org/10.1371/journal.pone.0084769.

(88) Kufareva, I.; Abagyan, R. Methods of Protein Structure Comparison; 2011; pp 231–257. https://doi.org/10.1007/978-1-61779-588-6_10.

(89) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(90) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol* **1993**, *234* (3), 779–815. https://doi.org/10.1006/jmbi.1993.1626.

(91) Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Šali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu Rev Biophys Biomol Struct* **2000**, *29* (1), 291–325. https://doi.org/10.1146/annurev.biophys.29.1.291.

(92) Alva, V.; Nam, S.-Z.; Söding, J.; Lupas, A. N. The MPI Bioinformatics Toolkit as an Integrative Platform for Advanced Protein Sequence and Structure Analysis. *Nucleic Acids Res* **2016**, *44* (W1), W410–W415. https://doi.org/10.1093/nar/gkw348.

(93) Bhachoo, J.; Beuming, T. Investigating Protein–Peptide Interactions Using the Schrödinger Computational Suite; 2017; pp 235–254. https://doi.org/10.1007/978-1-4939-6798-8_14.

(94) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(95) Hutter, M. C. Graph-Based Similarity Concepts in Virtual Screening. *Future Med Chem* **2011**, *3* (4), 485–501. https://doi.org/10.4155/fmc.11.3.

(96) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J Chem Inf Model* **2008**, *48* (1), 68–74. https://doi.org/10.1021/ci700286x.

(97) Vanii Jayaseelan, K.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural Product-Likeness Score Revisited: An Open-Source, Open-Data Implementation. *BMC Bioinformatics* **2012**, *13* (1), 106. https://doi.org/10.1186/1471-2105-13-106.

(98) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J Cheminform* **2016**, *8* (1), 61. https://doi.org/10.1186/s13321-016-0174-y.

(99) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. Bin; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J Nat Prod* **2021**, *84* (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.

(100) Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T. The Application of in Silico Drug-Likeness Predictions in Pharmaceutical Research. *Adv Drug Deliv Rev* **2015**, *86*, 2–10. https://doi.org/10.1016/j.addr.2015.01.009.

(101) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J Med Chem* **2008**, *51* (4), 817–834. https://doi.org/10.1021/jm701122q.

(102) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical Drug Properties Associated with in Vivo Toxicological Outcomes. *Bioorg Med Chem Lett* **2008**, *18* (17), 4872–4875. https://doi.org/10.1016/j.bmcl.2008.07.071.

(103) Shultz, M. D. Setting Expectations in Molecular Optimizations: Strengths and Limitations of Commonly Used Composite Parameters. *Bioorg Med Chem Lett* **2013**, *23* (21), 5980–5991. https://doi.org/10.1016/j.bmcl.2013.08.029.

(104) Shultz, M. D. Improving the Plausibility of Success with Inefficient Metrics. *ACS Med Chem Lett* **2014**, *5* (1), 2–5. https://doi.org/10.1021/ml4004638.

(105) Yusof, I.; Segall, M. D. Considering the Impact Drug-like Properties Have on the Chance of Success. *Drug Discov Today* **2013**, *18* (13–14), 659–666. https://doi.org/10.1016/j.drudis.2013.02.008.

(106) Beker, W.; Wołos, A.; Szymkuć, S.; Grzybowski, B. A. Minimal-Uncertainty Prediction of General Drug-Likeness Based on Bayesian Neural Networks. *Nat Mach Intell* **2020**, *2* (8), 457–465. https://doi.org/10.1038/s42256-020-0209-y.

(107) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat Chem* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

(108) Lee, K.; Jang, J.; Seo, S.; Lim, J.; Kim, W. Y. Drug-Likeness Scoring Based on Unsupervised Learning. *Chem Sci* **2022**, *13* (2), 554–565. https://doi.org/10.1039/D1SC05248A.

(109) Kavlock, R. J.; Ankley, G.; Blancato, J.; Breen, M.; Conolly, R.; Dix, D.; Houck, K.; Hubal, E.; Judson, R.; Rabinowitz, J.; Richard, A.; Setzer, R. W.; Shah, I.; Villeneuve, D.; Weber, E. Computational Toxicology—A State of the Science Mini Review. *Toxicological Sciences* **2008**, *103* (1), 14–27. https://doi.org/10.1093/toxsci/kfm297.

(110) EPA, U. S. A Framework for a Computational Toxicology Research Program. *Washington, DC: United States Environmental Protection Agency (EPA600/R-03/65)* **2003**.

(111) Maertens, A. Probabilistic Risk Assessment – the Keystone for the Future of Toxicology. *ALTEX* **2022**, *39* (1), 3–29. https://doi.org/10.14573/altex.2201081.

(112) Gallegos Saliner A., B. S. P. Mini-Review on Chemical Similarity and Prediction of Toxicity. *Current Computer Aided-Drug Design* **2006**, *2* (2), 105–122. https://doi.org/10.2174/157340906777441681.

(113) Organisation for Economic Co-operation and Development. *Grouping of Chemicals: Chemical Categories and Read-Across*. https://www.oecd.org/chemicalsafety/risk-assessment/groupingofchemicalschemicalcategoriesandread-across.htm.

(114) Chatterjee, M.; Roy, K. Chemical Similarity and Machine Learning-Based Approaches for the Prediction of Aquatic Toxicity of Binary and Multicomponent Pharmaceutical and Pesticide Mixtures against Aliivibrio Fischeri. *Chemosphere* **2022**, *308*, 136463. https://doi.org/10.1016/j.chemosphere.2022.136463.

(115) de Athayde, A. E.; de Araujo, C. E. S.; Sandjo, L. P.; Biavatti, M. W. Metabolomic Analysis among Ten Traditional “Arnica” (Asteraceae) from Brazil. *J Ethnopharmacol* **2021**, *265*, 113149. https://doi.org/10.1016/j.jep.2020.113149.

(116) Skinnider, M. A.; Dejong, C. A.; Franczak, B. C.; McNicholas, P. D.; Magarvey, N. A. Comparative Analysis of Chemical Similarity Methods for Modular Natural Products with a Hypothetical Structure Enumeration Algorithm. *J Cheminform* **2017**, *9* (1), 46. https://doi.org/10.1186/s13321-017-0234-y.

(117) Wei, B.; Hu, G.-A.; Zhou, Z.-Y.; Yu, W.-C.; Du, A.-Q.; Yang, C.-L.; Yu, Y.-L.; Chen, J.-W.; Zhang, H.-W.; Wu, Q.; Xuan, Q.; Xu, X.-W.; Wang, H. Global Analysis of the Biosynthetic Chemical Space of Marine Prokaryotes. *Microbiome* **2023**, *11* (1), 144. https://doi.org/10.1186/s40168-023-01573-3.

(118) Sánchez-Ruiz, A.; Colmenarejo, G. Systematic Analysis and Prediction of the Target Space of Bioactive Food Compounds: Filling the Chemobiological Gaps. *J Chem Inf Model* **2022**, *62* (16), 3734–3751. https://doi.org/10.1021/acs.jcim.2c00888.

(119) Rahman, Md. M.; Vadrev, S. M.; Magana-Mora, A.; Levman, J.; Soufan, O. A Novel Graph Mining Approach to Predict and Evaluate Food-Drug Interactions. *Sci Rep* **2022**, *12* (1), 1061. https://doi.org/10.1038/s41598-022-05132-y.

(120) 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.

(121) Ellin, N. R.; Miranda-Quintana, R. A.; Prentice, B. M. Extended Similarity Methods for Efficient Data Mining in Imaging Mass Spectrometry. *bioRxiv* **2023**. https://doi.org/10.1101/2023.07.27.550838.

(122) Stumpfe, D.; Bajorath, J. Methods for SAR Visualization. *RSC Adv.* **2012**, *2* (2), 369–378. https://doi.org/10.1039/C1RA00924A.

(123) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J Med Chem* **2007**, *50* (24), 5926–5937. https://doi.org/10.1021/jm070845m.

(124) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J Chem Inf Model* **2007**, *47* (1), 47–58. https://doi.org/10.1021/ci600338x.

(125) Yoshimori, A.; Tanoue, T.; Bajorath, J. Integrating the Structure–Activity Relationship Matrix Method with Molecular Grid Maps and Activity Landscape Models for Medicinal Chemistry

Applications. *ACS Omega* **2019**, *4* (4), 7061–7069. https://doi.org/10.1021/acsomega.9b00595.

(126) 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* (8), 795–802. https://doi.org/10.1007/s10822-014-9760-0.

(127) Naveja, J. J.; Medina-Franco, J. L. Finding Constellations in Chemical Space Through Core Analysis. *Front Chem* **2019**, *7*. https://doi.org/10.3389/fchem.2019.00510.

(128) Medina-Franco, J. L. Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *J Chem Inf Model* **2012**, *52* (10), 2485–2493. https://doi.org/10.1021/ci300362x.

(129) Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. Structure–Activity Relationships of Benzimidazole Derivatives as Antiparasitic Agents: Dual Activity-Difference (DAD) Maps. *Med. Chem. Commun.* **2011**, *2* (1), 44–49. https://doi.org/10.1039/C0MD00159G.

(130) Naveja, J. J.; Medina-Franco, J. L. ChemMaps: Towards an Approach for Visualizing the Chemical Space Based on Adaptive Satellite Compounds. *F1000Res* **2017**, *6*, 1134. https://doi.org/10.12688/f1000research.12095.2.

(131) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J Chem Inf Model* **2015**, *55* (6), 1136–1147. https://doi.org/10.1021/acs.jcim.5b00175.

(132) Dimova, D.; Stumpfe, D.; Bajorath, J. Quantifying the Fingerprint Descriptor Dependence of Structure–Activity Relationship Information on a Large Scale. *J Chem Inf Model* **2013**, *53* (9), 2275–2281. https://doi.org/10.1021/ci4004078.

(133) Gupta, S.; Aires-de-Sousa, J. Comparing the Chemical Spaces of Metabolites and Available Chemicals: Models of Metabolite-Likeness. *Mol Divers* **2007**, *11* (1), 23–36. https://doi.org/10.1007/s11030-006-9054-0.

(134) López-López, E.; Cerda-García-Rojas, C. M.; Medina-Franco, J. L. Tubulin Inhibitors: A Chemoinformatic Analysis Using Cell-Based Data. *Molecules* **2021**, *26* (9), 2483. https://doi.org/10.3390/molecules26092483.

(135) Khomtchouk, B. B.; Hennessy, J. R.; Wahlestedt, C. Shinyheatmap: Ultra Fast Low Memory Heatmap Web Interface for Big Data Genomics. *PLoS One* **2017**, *12* (5), e0176334. https://doi.org/10.1371/journal.pone.0176334.

(136) 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 Cheminform* **2016**, *8* (1), 63. https://doi.org/10.1186/s13321-016-0176-9.

(137) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J Am Chem Soc* **2013**, *135* (19), 7296–7303. https://doi.org/10.1021/ja401184g.

(138) Medina-Franco, J. L.; Naveja, J. J.; López-López, E. Reaching for the Bright StARs in Chemical Space. *Drug Discov Today* **2019**, *24* (11), 2162–2169. https://doi.org/10.1016/j.drudis.2019.09.013.

(139) Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol Inform* **2022**, *41* (11), 2200116. https://doi.org/10.1002/minf.202200116.

(140) López-López, E.; Medina-Franco, J. L. Towards Decoding Hepatotoxicity of Approved Drugs through Navigation of Multiverse and Consensus Chemical Spaces. *Biomolecules* **2023**, *13* (1), 176. https://doi.org/10.3390/biom13010176.

(141) Wang, B.; Mezlini, A. M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat Methods* **2014**, *11* (3), 333–337. https://doi.org/10.1038/nmeth.2810.

(142) Weston, J.; Elisseeff, A.; Zhou, D.; Leslie, C. S.; Noble, W. S. Protein Ranking: From Local to Global Structure in the Protein Similarity Network. *Proceedings of the National Academy of Sciences* **2004**, *101* (17), 6559–6563. https://doi.org/10.1073/pnas.0308067101.

(143) 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* **2022**, *62* (9), 2186–2201. https://doi.org/10.1021/acs.jcim.1c01013.

(144) Lo, Y.-C.; Senese, S.; Damoiseaux, R.; Torres, J. Z. 3D Chemical Similarity Networks for Structure-Based Target Prediction and Scaffold Hopping. *ACS Chem Biol* **2016**, *11* (8), 2244–2253. https://doi.org/10.1021/acschembio.6b00253.

(145) López-López, E.; Barrientos-Salcedo, C.; Prieto-Martínez, F. D.; Medina-Franco, J. L. In Silico Tools to Study Molecular Targets of Neglected Diseases: Inhibition of TcSir2rp3, an Epigenetic Enzyme of Trypanosoma Cruzi; 2020; pp 203–229. https://doi.org/10.1016/bs.apcsb.2020.04.001.

(146) Naveja, J. J.; Norinder, U.; Mucs, D.; López-López, E.; Medina-Franco, J. L. Chemical Space, Diversity and Activity Landscape Analysis of Estrogen Receptor Binders. *RSC Adv* **2018**, *8* (67), 38229–38237. https://doi.org/10.1039/C8RA07604A.

(147) Gaytán-Hernández, D.; Chávez-Hernández, A. L.; López-López, E.; Miranda-Salas, J.; Saldívar-González, F. I.; Medina-Franco, J. L. Art Driven by Visual Representations of Chemical Space. *J Cheminform* **2023**, *15* (1), 100. https://doi.org/10.1186/s13321-023-00770-4.

(148) Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J Cheminform* **2021**, *13* (1), 33. https://doi.org/10.1186/s13321-021-00504-4.

(149) Kuo, C.; Glover, F.; Dhir, K. S. Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming*. *Decision Sciences* **1993**, *24* (6), 1171–1185. https://doi.org/10.1111/j.1540-5915.1993.tb00509.x.

(150) Verhellen, J. Graph-Based Molecular Pareto Optimisation. *Chem Sci* **2022**, *13* (25), 7526–7535. https://doi.org/10.1039/D2SC00821A.

(151) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Physical Chemistry Chemical Physics* **2022**, *24* (1), 444–451. https://doi.org/10.1039/D1CP04019G.

(152) Vogt, M.; Stumpfe, D.; Maggiora, G. M.; Bajorath, J. Lessons Learned from the Design of Chemical Space Networks and Opportunities for New Applications. *J Comput Aided Mol Des* **2016**, *30* (3), 191–208. https://doi.org/10.1007/s10822-016-9906-3.

(153) Flores-Padilla, E. A.; Juárez-Mercado, K. E.; Naveja, J. J.; Kim, T. D.; Alain Miranda-Quintana, R.; Medina-Franco, J. L. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. *Mol Inform* **2022**, *41* (6), 2100285. https://doi.org/10.1002/minf.202100285.

(154) 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. https://doi.org/10.1021/acs.jcim.3c00520.

(155) Bajusz, D.; Miranda-Quintana, R. A.; Rácz, A.; Héberger, K. Extended Many-Item Similarity Indices for Sets of Nucleotide and Protein Sequences. *Comput Struct Biotechnol J* **2021**, *19*, 3628–3639. https://doi.org/10.1016/j.csbj.2021.06.021.

(156) Rácz, A.; Dunn, T. B.; Bajusz, D.; Kim, T. D.; Miranda-Quintana, R. A.; Héberger, K. Extended Continuous Similarity Indices: Theory and Application for QSAR Descriptor Selection. *J Comput Aided Mol Des* **2022**, *36* (3), 157–173. https://doi.org/10.1007/s10822-022-00444-7.

(157) Danishuddin; Khan, A. U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discov Today* **2016**, *21* (8), 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013.

(158) Redžepović, I.; Furtula, B. Chemical Similarity of Molecules with Physiological Response. *Mol Divers* **2023**, *27* (4), 1603–1612. https://doi.org/10.1007/s11030-022-10514-5.

(159) López-Pérez, K.; López-López, E.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Sampling and Mapping Chemical Space with Extended Similarity Indices. *Molecules* **2023**, *28* (17), 6333. https://doi.org/10.3390/molecules28176333.

(160) Rácz, A.; Mihalovits, L. M.; Bajusz, D.; Héberger, K.; Miranda-Quintana, R. A. Molecular Dynamics Simulations and Diversity Selection by Extended Continuous Similarity Indices. *J Chem Inf Model* **2022**, *62* (14), 3415–3425. https://doi.org/10.1021/acs.jcim.2c00433.

(161) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *J Chem Phys* **2019**, *151* (7). https://doi.org/10.1063/1.5109531.

(162) Theobald, D. L.; Wuttke, D. S. THESEUS: Maximum Likelihood Superpositioning and Analysis of Macromolecular Structures. *Bioinformatics* **2006**, *22* (17), 2171–2172. https://doi.org/10.1093/bioinformatics/btl332.

(163) Theobald, D. L.; Wuttke, D. S. Accurate Structural Correlations from Maximum Likelihood Superpositions. *PLoS Comput Biol* **2008**, *4* (2), e43. https://doi.org/10.1371/journal.pcbi.0040043.

(164) Tribello, G. A.; Gasparotto, P. Using Dimensionality Reduction to Analyze Protein Trajectories. *Front Mol Biosci* **2019**, *6*. https://doi.org/10.3389/fmolb.2019.00046.

(165) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J Chem Inf Model* **2014**, *54* (9), 2555–2561. https://doi.org/10.1021/ci500319f.

(166) Bouysset, C.; Fiorucci, S. ProLIF: A Library to Encode Molecular Interactions as Fingerprints. *J Cheminform* **2021**, *13* (1), 72. https://doi.org/10.1186/s13321-021-00548-6.

(167) Yasuo, N.; Sekijima, M. Improved Method of Structure-Based Virtual Screening via Interaction-Energy-Based Learning. *J Chem Inf Model* **2019**, *59* (3), 1050–1061. https://doi.org/10.1021/acs.jcim.8b00673.

(168) Trozzi, F.; Wang, X.; Tao, P. UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study. *J Phys Chem B* **2021**, *125* (19), 5022–5034. https://doi.org/10.1021/acs.jpcb.1c02081.

(169) Pearson, K. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **1901**, *2* (11), 559–572.

(170) Molgedey, L.; Schuster, H. G. Separation of a Mixture of Independent Signals Using Time Delayed Correlations. *Phys Rev Lett* **1994**, *72* (23), 3634–3637. https://doi.org/10.1103/PhysRevLett.72.3634.

(171) Doerr, S.; Ariz-Extreme, I.; Harvey, M. J.; Fabritiis, G. De. Dimensionality Reduction Methods for Molecular Simulations. 2017.

(172) Huang, H.; Wang, Y.; Rudin, C.; Browne, E. P. Towards a Comprehensive Evaluation of Dimension Reduction Methods for Transcriptomic Data Visualization. *Commun Biol* **2022**, *5* (1), 719. https://doi.org/10.1038/s42003-022-03628-x.

(173) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem Rev* **2021**, *121* (16), 9722–9758. https://doi.org/10.1021/acs.chemrev.0c01195.

(174) Tournier, A. L.; Smith, J. C. Principal Components of the Protein Dynamical Transition. *Phys Rev Lett* **2003**, *91* (20), 208106. https://doi.org/10.1103/PhysRevLett.91.208106.

(175) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Structure, Function, and Bioinformatics* **2005**, *58* (1), 45–52. https://doi.org/10.1002/prot.20310.

(176) Oide, M.; Sugita, Y. Protein Folding Intermediates on the Dimensionality Reduced Landscape with UMAP and Native Contact Likelihood. *J Chem Phys* **2022**, *157* (7). https://doi.org/10.1063/5.0099094.

(177) Sittel, F.; Stock, G. Perspective: Identification of Collective Variables and Metastable States of Protein Dynamics. *J Chem Phys* **2018**, *149* (15). https://doi.org/10.1063/1.5049637.

(178) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput* **2007**, *3* (6), 2312–2334. https://doi.org/10.1021/ct700119m.

(179) Torda, A. E.; van Gunsteren, W. F. Algorithms for Clustering Molecular Dynamics Configurations. *J Comput Chem* **1994**, *15* (12), 1331–1340. https://doi.org/10.1002/jcc.540151203.

(180) Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* **2015**, *2* (2), 165–193. https://doi.org/10.1007/s40745-015-0040-1.

(181) Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Trans Knowl Discov Data* **2015**, *10* (1), 1–51. https://doi.org/10.1145/2733381.

(182) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; others. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *kdd*; 1996; Vol. 96, pp 226–231.

(183) Gholizadeh, N.; Saadatfar, H.; Hanafi, N. K-DBSCAN: An Improved DBSCAN Algorithm for Big Data. *J Supercomput* **2021**, *77* (6), 6214–6235. https://doi.org/10.1007/s11227-020-03524-3.

(184) Hunkler, S.; Diederichs, K.; Kukharenko, O.; Peter, C. Fast Conformational Clustering of Extensive Molecular Dynamics Simulation Data. *J Chem Phys* **2023**, *158* (14). https://doi.org/10.1063/5.0142797.