Cheminformatics Tools to Explore the Chemical Space of Peptides and Natural Products

Inaugural dissertation

of the Faculty of Science,

University of Bern

presented by

Alice Capecchi

from La Rotta, PI, Italy

Supervisor of the doctoral thesis:

Prof. Dr. Jean-Louis Reymond

Department of Chemistry and Biochemistry

Cheminformatics Tools to Explore the Chemical Space of Peptides and Natural Products

Inaugural dissertation

of the Faculty of Science,

University of Bern

presented by

Alice Capecchi

from La Rotta, PI, Italy

Supervisor of the doctoral thesis:

Prof. Dr. Jean-Louis Reymond
Department of Chemistry and Biochemistry

Accepted by the Faculty of Science.

Bern, 8th of October, 2021

The Dean

Prof. Dr. Zoltan Balogh

Acknowledgments

Many thanks to my supervisor Jean-Louis for the advice, the directions, and the constructive arguments.

Many thanks to Daniel, Sacha, Josep, and Mahendra, who patiently taught me a lot, and to all my colleagues for the brainstorming, the beers, and the parties. Many became close friends, and thanks to them, I will bring with me countless fun memories of these four years.

Many thanks to my parents Sergio and Antonella, to my brother Gianni, and to my childhood friends Giulia and Serena, whom I can always count on even from a distance. *Mi mancate di continuo e vorrei vedervi piú spesso.*

Many thanks to my boyfriend Thomas, who has been my rock in these moments of change and emotional chaos.

Abstract

Cheminformatics facilitates the analysis, storage, and collection of large quantities of chemical data, such as molecular structures and molecules' properties and biological activity, and it has revolutionized medicinal chemistry for small molecules. However, its application to larger molecules is still underrepresented. This thesis work attempts to fill this gap and extend the cheminformatics approach towards large molecules and peptides.

This thesis is divided into two parts. The first part presents the implementation and application of two new molecular descriptors: macromolecule extended atom pair fingerprint (MXFP) and MinHashed atom pair fingerprint of radius 2 (MAP4). MXFP is an atom pair fingerprint suitable for large molecules, and here, it is used to explore the chemical space of non-Lipinski molecules within the widely used PubChem and ChEMBL databases. MAP4 is a MinHashed hybrid of substructure and atom pair fingerprints suitable for encoding small and large molecules. MAP4 is first benchmarked against commonly used atom pairs and substructure fingerprints, and then it is used to investigate the chemical space of microbial and plants natural products with the aid of machine learning and chemical space mapping.

The second part of the thesis focuses on peptides, and it is introduced by a review chapter on approaches to discover novel peptide structures and describing the known peptide chemical space. Then, a genetic algorithm that uses MXFP in its fitness function is described and challenged to generate peptide analogs of peptidic or non-peptidic queries. Finally, supervised and unsupervised machine learning is used to generate novel antimicrobial and non-hemolytic peptide sequences.

Tables of Contents

Thesis Scope & Outline

While cheminformatics has revolutionized medicinal chemistry for small molecules, its application to larger molecules still lacks representation. For instance, macromolecules can be the object of rigorous molecular dynamics and conformational studies, but an orthogonal two-dimensional approach that allows a fast analysis and visualization of their space is still missing.

In fact, in the context of their two-dimensional structural representation, macromolecules such as peptides and glycans are almost exclusively handled by bioinformatics using sequences formed by specified building blocks.[1] This encoding is well established and successful; however, it has limitations. In fact, assessing similarity works only within a specific macromolecule class and within the established building blocks and topology frame. For instance, a classical sequence-based bioinformatics approach would not allow encoding modified peptides such as vancomycin,[2] or peptides with unusual topologies, such as the antimicrobial peptide dendrimer **G3KL**.[3] These molecules can instead be described with a cheminformatics approach, for instance, by using molecular fingerprints. However, cheminformatics has primarily focused on small molecules and presents its limitation in larger molecules encoding.

Prior to the work reported in this thesis, studies on using a two-dimensional molecular fingerprint for peptides[4–6] have been carried out in the Reymond group. This thesis' scope is to expand these initial results by building a solid cheminformatic toolbox for large molecules, including molecular fingerprints, chemical space visualization, and generation of novel peptide sequences. The contributions made towards this objective have been published as scientific articles. In this thesis, they are collected and divided into two main sections: (i) encoding of large molecules and visualization of their chemical space and (ii) peptide chemical space and

generation of novel peptide sequences. Each section has an introductory chapter, namely chapter one and chapter six.

Part one – Encoding of large molecules and visualization of their chemical space

- **Chapter One.** This chapter consists of a general introduction to cheminformatics, molecular fingerprints, and chemical space visualization. Since they are important for the understanding of the following chapters, the extended connectivity fingerprints (ECFPs),[7] MinHashed fingerprints (MHFPs),[8] extended atom pair fingerprint (Xfp),[9] RDKit atom pair fingerprint (RDKit AP),[10] similarity maps,[11] and Tree Maps (TMAP)[12] are treated in greater detail. Then, the state of the art and limitations of molecular fingerprints and chemical space visualization for larger molecules and peptides are summarized.
- **Chapter Two.** An adaptation of Xfp for large molecules is introduced: the macromolecule extended atom pair fingerprint (MXFP). Then, the molecules within PubChem and ChEMBL databases breaking more than one of Lipinski's rules of five[13] are encoded with MXFP, and their similarity maps are visualized. The resulting interactive maps allow for exploring a subset of widely used databases that contains often neglected larger structures and macromolecules. This chapter is based on the peer-reviewed publication "PubChem and CHEMBL beyond Lipinski".[14]
- **Chapter Three.** A novel descriptor for both small and large molecules is described: the MinHashed atom pair fingerprint of radius 2 (MAP4). MAP4 is a MinHashed fingerprint that bridges between substructure fingerprints such as ECFPs and MHFPs and atom pair fingerprints such as the RDKit atom pair fingerprint. Then, MAP4 is benchmarked against MHFP6, ECFP4, the topological torsion fingerprint (TT),[15] and RDKit AP using a peptide extended version of the Riniker and Landrum small molecules fingerprint benchmark.[16] Contrarily to the other tested fingerprint, MAP4 scores excellent performances with both small molecules and peptides. Furthermore, MAP4 is shown to encode the human

metabolome database correctly.[17] This chapter is based on the peer-reviewed publication "One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome".[18]

- **Chapters Four and Five.** Due to its capabilities in encoding molecules covering a vast size range, MAP4 is suitable to represent and visualize natural products. In chapter four, the analysis of the microbial natural products using the Natural product Atlas[19] is carried over, and the observation that both a MAP4 TMAP of this space and a MAP4 SVM classifier can distinguish the natural products based on their fungal or bacterial origin is made. In chapter five, the analysis is extended to natural products from plants using the recently published Collection of open natural products (COCONUT)[20] and obtaining similar results. In an unprecedented way, this analysis allows for an interactive visualization of the natural products' chemical space and the origin assignment of a natural product using a fingerprint-based classifier. Chapter four is based on the peer-reviewed publication "Assigning the origin of microbial natural products by chemical space map and machine learning".[21] Chapter five is based on the ChemRxiv preprint "Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database".[22]

Part two – Peptide chemical space and generation of novel peptide sequences

- **Review Chapter Six.** The main computational approaches to the generation of novel peptides are listed and the state of the art summarized. Then, the known chemical space of linear peptides up to 50 residues collected from the available databases is encoded using MAP4 and visualized with TMAP. Review chapter six is based on the peer-reviewed publication "Peptides in Chemical Space".[23]
- **Chapter Seven.** A random subset of the virtual chemical space of linear, cyclic, and dendritic peptides is sampled. Then, a novel methodology to translate the dendritic and cyclic peptide sequences into SMILES is described and used. The resulting SMILES are encoded with MXFP, a PCA is applied, and the first three principal components are

visualized using the interactive visualization tool Faerun[24]. The implementation of the peptide design genetic algorithm (PDGA), a genetic algorithm that generates peptides sequences and uses MXFP in its fitness function, is described. Next, PDGA is successfully challenged to produce peptide analogs of peptidic and non-peptidic queries. PDGA allows for the exploration of specific areas of the peptide chemical space through analogs generation. Although its first implementation uses MXFP, PDGA has been adapted to use the RDKit AP fingerprint and MAP4. This adaptation source code is publicly available at https://github.com/reymond-group/PDGA-MAP4_AP. Chapter seven is based on the peer-reviewed publication "Populating chemical space with peptides using a genetic algorithm".[25]

- **Chapter Eight.** Machine learning is challenged with the design of novel non-hemolytic antimicrobial peptides (AMPs) against three problematic and often drug-resistant pathogens: P. aeruginosa, A. baumannii, and S. aureus. Highly curated hemolysis and antimicrobial activity data on linear peptides are used to train a generative model, fine-tune it to generate non-hemolytic peptides against specific bacterial strains, and train an antimicrobial activity and a hemolysis classifier. Out of 28 synthesized peptides, eight are non-hemolytic and active against the bacterial strains used in the sequence design. The presented work permits the exploration of peptide subspaces characterized by a determined activity. The used workflow can be adapted to different antimicrobial activities using different subsets during the fine-tuning of the generative model. Furthermore, the different recurrent neural networks models can be used separately. Chapter eight is based on the peer-reviewed publication "Machine Learning Designs Non-Hemolytic Antimicrobial Peptides".[26]

Chapter nine. Conclusive thoughts on the impact of this thesis in the field are followed by a brief introduction to possible future work within the framework of the presented projects.

List of Publications

First author articles have been incorporated into this thesis as separate chapters, and here they are listed from the more recent to the least recent with the corresponding chapter number and DOI.

Capecchi Alice and Reymond Jean-Louis. Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database. *ChemRxiv preprint*, 2021. Chapter Five. https://doi.org/10.33774/chemrxiv-2021-gxjgc

Capecchi Alice*, Cai Xingguang*, Personne Hippolyte, Kohler Thilo, van Delden Christian, and Reymond Jean-Louis. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chemical Science*, 2021. *These authors contributed equally. Chapter Eight. https://doi.org/10.1039/D1SC01713F

Capecchi Alice and Reymond Jean-Louis. Peptides in chemical space. *Medicine in Drug Discovery*, 2021. Review chapter six. https://doi.org/10.1016/j.medidd.2021.100081

Capecchi Alice and Reymond, Jean-Louis. Assigning the origin of microbial natural products by chemical space map and machine learning. *Biomolecules*, 2020. Chapter Four. https://doi.org/10.3390/biom10101385

Capecchi Alice, Probst Daniel, and Reymond Jean-Louis. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 2020. Chapter Three. https://doi.org/10.1186/s13321-020-00445-4

Capecchi Alice, Zhang Alain, and Reymond Jean-Louis. *Journal of chemical information and modeling*, 2019. Chapter Seven. https://doi.org/10.1021/acs.jcim.9b01014

Capecchi Alice, Awale Mahendra, Probst Daniel, and Reymond Jean-Louis. PubChem and CHEMBL beyond Lipinski. *Molecular informatics*, 2019. Chapter Two. https://doi.org/10.1002/minf.201900016

The following articles were co-authored and are not incorporated into this thesis.

- Siriwardena Thissa, Capecchi Alice, Gan Bee-Ha, Jin Xian, He Runze, Wei Dengwen, Ma Lan, Köhler Thilo, Van Delden Christian, Javor Sacha, and Reymond Jean-Louis. Optimizing antimicrobial peptide dendrimers in chemical space. *Angewandte Chemie*, 2018. https://doi.org/10.1002/anie.201802837
- Di Bonaventura Ivan, Baeriswyl Stéphane, Capecchi Alice, Gan Bee-Ha, Jin, Xian, Siriwardena Thissa, He Runze, Köhler Thilo, Pompilio Arianna, Di Bonaventura Giovanni, Van Delden Christian, Javor Sacha, and Reymond Jean-Louis. An antimicrobial bicyclic peptide from chemical space against multidrug resistant Gram-negative bacteria. *Chemical communications*, 2018. https://doi.org/10.1039/C8CC02412J

Part One – Encoding of Large Molecules and Visualization of their Chemical Space

Chapter One – General Introduction

Cheminformatics facilitates the analysis, storage, and collection of large quantities of chemical data, such as molecular structures and molecules' properties and biological activity. While several concepts behind the discipline are much older,[27,28] the term cheminformatics was first defined by Frank Brown in 1998[29] as an abbreviation of "chemical informatics". The need for this definition was probably a consequence of the dramatic importance that cheminformatics acquired within the drug discovery process after the advent of high throughput drug screening and parallel synthesis between the 1980s and 1990s.[30] In fact, this increased capabilities of the chemical field produced an unprecedented growth of the synthesized compounds creating the need for tools to handle them.[1] Need that cheminformatics and, more recently, machine learning have transformed into an opportunity to exploit the information within the exponentially growing chemical data.[31] Nowadays, cheminformatics is deeply interconnected with the drug discovery process: from screening protein databases for possible targets during the target identification phase, to virtual screening, database generations, property predictions in the lead discovery and optimization process, and ADMET profiling prediction prior to preclinical studies.[32–35] Two typical cheminformatics approaches are molecular fingerprints encoding and chemical space exploration. The following sections summarize their state of the art and the limitations of their application to large molecules and peptides.

1.1 Molecular Descriptors

A frequent task in a cheminformatics project is the comparison between two molecules and their similarity assessment. This task is typically tackled using molecular descriptors.

Molecular descriptors can be roughly classified as 1, 2, or 3 dimensional (D). 1D descriptors take into account the physicochemical properties of a molecule and the presence of specific substructures. 2D and 3D descriptors consider the whole molecule. 3D descriptors bring the highest level of information. However, they require 3-dimensional information of a molecule, which is rarely experimentally available, and its calculation with molecular docking and dynamics is often slow to compute. 2D descriptors contain information on the complete 2-D structure of molecules, and even without the complexity of 3D representation, their performance is generally comparable to the one of 3D descriptors for most typical drug discovery tasks.[36,37] The following sections will focus on two types of 2D descriptors that are important to understand the following chapters of this thesis: circular substructure fingerprints and atom pair fingerprints.

1.1.1 Circular Substructure Fingerprints

Circular substructure fingerprints encode each atom in a molecule with its circular substructure up to a defined radius. A notorious example of this class of fingerprints are the extended-connectivity fingerprints (ECFPs).[7] Recently, Daniel Probst and Jean-Louis Reymond published the MinHashed fingerprints (MHFPs), a variation of ECFPs which uses methods from natural language processing and data mining.[8] While circular substructure fingerprints are well-performing with small molecules,[16,18,38] they are less suitable for large and repetitive structures. In fact, to avoid an overly specific encoding, generally, the radius of the considered substructure is relatively small, and when a molecule is large, substructure fingerprints can lose the perception of the whole structure.[18] Since they are important for the understanding of the MinHashed atom pair fingerprint (MAP4),[18] which is introduced in the second chapter of this thesis, the RDKit implementation of ECFPs[7] (Figure 1.2a) and the implementation of MHFPs (Figure 1.2b) are described in detail in the following sections.

ECFPs. ECFPs (Figure 1a) are among the most widely used and best-performing fingerprints available in cheminformatics. In the RDKit implementation of the ECFP fingerprints, all atoms are assigned hashed numerical identifiers that encode the properties of the atom itself and the properties of their neighboring atoms within their circular environment of the given radius. The considered atom properties are the atom's element, the number of heavy neighbors, number of hydrogens, charge, isotope, and if the atom belongs to a ring. The unique numerical identifiers representing the molecule are then brought to a fixed-length binary array using the modulo operation. Following this logic, ECFP0 will encode each atom in a molecule using solely the information related to the atom itself, ECFP2 will encode each atom with its properties and the properties of the atoms in its circular environment of diameter 2, and so on. The resulting fingerprint is binary, and its default version has 1024 bits. The most used ECFPs are the ones of diameter 4 (ECFP4) and diameter 6 (ECFP6).

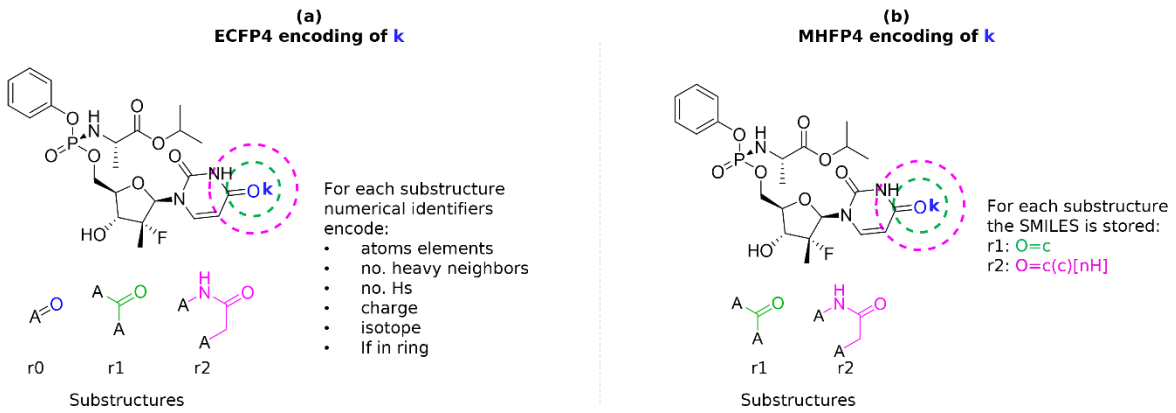


Figure 1. Substructure fingerprints encoding schematic. (a) ECFP4 encoding of atom **k**. (b) MHFP4 encoding of atom **k**.

MHFPs. MHFPs (Figure 1b) is a family of very recently published fingerprints that, as ECFPs, encode each atom in a molecule with its circular environment up to a given diameter. However, MHFPs represent atom's properties and environment using their Simplified molecular-input line-entry systems (SMILES). David Weininger first introduced SMILES in 1988.[39] They exemplify a molecular graph as a string, where atoms are represented with their standard

abbreviation (e.g., C for carbon atoms) and ramifications with brackets. In addition, bonds, stereochemistry, charges, and aromaticity are represented with different characters. In MHFPs, the substructures SMILES up to the defined diameter are extracted for each atom in a molecule, and then they are made unique and mapped to numbers using the SHA-1 hashing function.[40] Finally, the set of unique numbers is MinHashed[41] to a fixed length. MinHashing is a relatively slow procedure compared to the application of the modulo operation. However, the Jaccard similarity (Equation 1) of two MinHashed fingerprints *a* and *b* can be estimated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the fingerprint length.

$$J(A,B) = \frac{A \cap B}{A \cup B} \quad \text{Equation 1}$$

Therefore, using a MinHashed fingerprint has the advantage of allowing for fast similarity calculations. Furthermore, the fingerprint values derived from the MinHashing procedure can be directly stored within LSH forest trees,[8,42] making the methodology particularly compatible with the TMAP visualization described later in this introduction. The default MHFP has radius 3 and 1024 dimensions.

1.1.2 Atom Pair Fingerprints

Atom pair fingerprints represent molecular structures through atom pairs instead of single atoms. They encode atom pairs as the properties of both atoms and the distance that separates them. Depending on the maximum encoded distance between atoms, atom pair fingerprints can be used for small molecules or large molecules. For instance, the chemically advanced template search (CATS)[43] and Xfp[9] consider atom distances only up to 10 bonds, while the RDKit Atom pair fingerprint (RDKit AP)[10] encodes unlimited atom distances. Another 2D atom pair fingerprint suitable for large molecules is the 2D-Protein fingerprint (2DP).[6] 2DP considers long atom distances and has been proven capable of comparing non-linear peptides;[4–6]

however, 2DP assigns all atoms to the α-carbon atom of their parent amino acid residue, and then it counts the distances between the pairs of α-carbon atoms. Therefore, its use is limited to peptides. Atom pair fingerprints have been shown to represent molecular shape and pharmacophore, and they have been used for scaffold hopping.[9,43] However, they can represent fewer details than substructure fingerprints, which remain the fingerprints of choice for most cheminformatics projects involving small molecules. In the next section, we will focus on the implementation of the RDKit AP (Figure 2a) and Xfp (Figure 2b) since they are important to understand the macromolecule extended atom pair fingerprint (MXFP)[14] and the MAP4 fingerprint discussed in the first two chapters of this thesis.

RDKit AP. In the RDKit Atom pair fingerprint, all atom pairs are encoded with a numerical identifier that collects the atom type of the two atoms that form them and the distance in bonds that separates them. The considered properties to determine the atom type are the atom's element, number of heavy neighbors, and number of pi electrons. Since it encodes substructural paths (*i.e.*, the atom pair and the bond path between them), this fingerprint can be further labeled as path-based atom pair or substructural atom pair. The RDKit AP can be brought to fixed-length using the modulo operation, resulting in a binary array that by default has 2048 bits.

Xfp. Xfp is an atom pair fingerprint developed to encode small molecules and optimized using the DUD[44] dataset. Xfp classifies each atom as belonging to one or more of the following categories: hydrophobic (Hyb), H-bond acceptor (HBA), H-bond donor (HBD), and planar atoms (sp2). Atom pairs from bond distance one up to bond distance ten are counted within all categories and across the HBA and HBD categories. To avoid an overrepresentation of molecular size, the values are then normalized by the number of atoms in each specific category, and for the HBA-HBD cross pair, the HBA atom number is used. Finally, the bit values in HBA same-property pairs are doubled to increase their weight, resulting in a 55

dimensions array. Since Xfp, as well as CATS and 2DP, counts the number of atom pairs at defined bond distances, it can be further labeled as counted atom pair fingerprint.

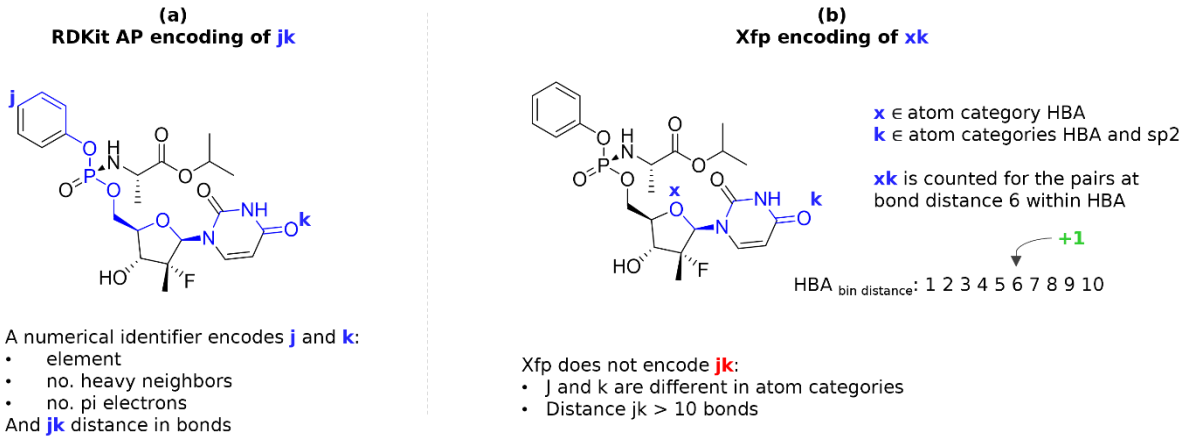


Figure 2. Atom pair fingerprints encoding schematic. (a) RDKit AP encoding of atom pair **jk**. (b) Xfp encoding of atom pair **xk**.

1.2 Chemical Space

The chemical space describes all possible molecules, including the available ones and those yet to be discovered.[45–47] This space can be considered limitless. However, Cheminformatics has been historically focused on the space of small drug-like molecules. This space has been estimated to be 10^{60} molecules by Bohacek,[48] 10^{33} molecules by Polishchuk,[49] and 10^{24} molecules by Peter Ertl.[50] Independently from the different size estimations, the complete enumeration of this space is not feasible, and the current most significant attempt is represented by the generated databases (GDBs).[51–53] However, generating, encoding, and visualizing the known chemical space, virtual subsets of the chemical space, or the chemical space of a specific set of molecules, has been successfully attempted, and it has contributed to drug discovery.[47,54,55]

1.2.1 Visualization of Chemical Spaces

As discussed in the previous section, the enumeration of the entire chemical space is not feasible. However, it is instead possible to focus on the characterization and visualization of the chemical space of a specific set of molecules. A chemical space is not only characterized by its molecules but also by the descriptor chosen to encode them. However, to effectively characterize molecules, molecular descriptors have more than three elements, while three are the maximum number of dimensionalities that are non-trivial to visualize. To overcome this impasse and visualize a chemical space, a dimensionality reduction methodology needs to be applied. Different methods are available, including t-distributed stochastic neighbor embedding (tSNE),[56] uniform manifold approximation and projection for dimension reduction (UMAP),[57] principal components analysis (PCA),[58] similarity maps,[11] and TMAPs.[12] Here, we will discuss the three approaches that will be found later in the chapters of this thesis.

PCA. A common approach to reducing the dimensionality of a dataset encoded with a multi-dimensional fingerprint is to do PCA and then visualize the first two or three principal components in a scatter plot.[59] PCA dates back to the beginning of the 19th century,[58] and it consists in reducing the dimensionality of a dataset while preserving as much information as possible. Preserving as much information as possible translates into finding new variables that are linear functions of those in the original dataset, maximize variance, and that are uncorrelated with each other.

Similarity maps. The visualization of the first two or three principal components of a PCA directly applied to a fingerprint encoding of a dataset does not always work well with molecular fingerprints that store a large amount of information, such as ECFPs or atom pair fingerprints. In fact, the two first principal components' explained variance can remain below 50% in these cases, making the visualization not informative. Similarity maps were developed in the Reymond group as an attempt to solve this problem.[11] In a similarity map, the PCA is computed

on the similarity matrix formed by the normalized similarity of each molecule in the database to a set of molecules (so-called satellites) randomly picked from the database. Similarity maps in the feature space of atom pairs and substructure fingerprints typically show a variance of the two first principal components above 50% and allow for a meaningful visualization.

Tree Map (TMAP). TMAP[12] layout can be calculated from a locality-sensitive hashing (LSH) forest[42] or a similarity matrix. *K* nearest neighbors (NNs) in the chosen feature space are extracted from the LSH forest or the similarity matrix to form a graph in which nodes are the structures and edges are the NN relationships weighted by the fingerprint distance. Kruskal's algorithm[60] is then applied to remove cycles and find the path with the lowest total distance between all the molecules in the graph. Finally, Fearun[24] can be used to display the obtained minimum spanning tree interactively. MHFPs are particularly suitable for this methodology since the indices generated by the MinHash procedure can be readily used to create an LSH forest.

1.2.2 Chemical Space of Large Molecules

More recently, drug discovery has extended its interest to the chemical space of larger molecules. For instance, the chemical space of the currently available natural products (NPs) contains molecules having a wide size range (Table 9 and Figure 46a), and it has been explored using physicochemical properties, the presence of specific substructures, hierarchical scaffold classification, and circular substructure fingerprints.[61–67] However, when looking at a more uniform and repetitive space such as the human metabolome[17] chemical space, substructure fingerprints become almost entirely incapable of distinguishing different entries (Figure 16 panels a, c, d, and e and Figure 15 panels b and c).[18] In fact, when talking about chemical space, one should not forget that this concept is deeply interconnected with a specific descriptor space

and that the organization of a chemical space can dramatically change when changing the descriptor used to encode the molecular structures.

For instance, since the MHFP6 fails to distinguish between related metabolites properly, the TMAP of the human metabolome database in this feature space is not very informative (Figure 15b). Furthermore, an analysis of the occupancy of fingerprint value bins illustrates how for MHFP6, ECFP4, and TT, the ten most populated fingerprint value bins contain many molecules, thousands for ECFP4 and MHFP6, and hundreds for TT (Figure 15c). On the contrary, the RDKit AP seems better equipped to distinguish between human metabolites, with only two or three molecules per fingerprint bin. However, as previously discussed, the RDKit AP, due to the lack of detailed substructural information, cannot distinguish between structurally correlated small molecules such as the metabolites HMDB0059800 and HMDB0059800 (Figure 16b). This analysis is found in its extended version in the second chapter of this thesis.

Due to their generally larger and repetitive structures, the same limitations can be extended to peptides and their chemical space, which can be explored using substructure fingerprints only when considering very short sequences.[68] In fact, TT, ECFP4, and MHFP6 are incapable of distinguishing scrambled peptide structures (Figure 16d). The known and the virtual chemical space of peptides, together with a summary of the methods applied to explore it, is thoroughly discussed in review chapter six. Review chapter six also contains general explanations of genetic algorithms and machine learning, which are introductory topics for the last two chapters of this thesis.

Chapter Two – PubChem and ChEMBL beyond Lipinski

This work is based on the peer-reviewed publication:

Capecchi, A.; Awale, M.; Probst, D.; Reymond, J.-L. PubChem and ChEMBL beyond Lipinski. Mol. Inform. 2019, 38 (5). https://doi.org/10.1002/minf.201900016.

This work is republished in this dissertation with the authorization of Wiley-VCH Verlag with Order License ID 1157319-1.

Abstract

Seven million of the currently 94 million entries in the PubChem database break at least one of the four Lipinski constraints for oral bioavailability, 183,185 of which are also found in the ChEMBL database. These non-Lipinski PubChem (NLP) and ChEMBL (NLC) subsets are interesting because they contain new modalities that can display biological properties not accessible to small molecule drugs. Unfortunately, the current search tools in PubChem and ChEMBL are designed for small molecules and are not well suited to explore these subsets, which therefore remain poorly appreciated. Herein we report MXFP (macromolecule extended atom-pair fingerprint), a 217-D fingerprint tailored to analyze large molecules in terms of molecular shape and pharmacophores. We implement MXFP in two web-based applications, the first one to visualize NLP and NLC interactively using Faerun (http://faerun.gdb.tools/), the second one to perform MXFP nearest neighbor searches in NLP. We show that these tools provide a meaningful insight into the diversity of large molecules in NLP and NLC.

2.1 Introduction

PubChem and ChEMBL are public repositories of molecules and their biological activity.[69,70] While both databases contain a vast majority of small molecules, they also contain a small percentage of larger biomolecules such as peptides, oligonucleotides, oligosaccharides, and large natural products, as well as synthetic macromolecules such as peptide nucleic acids, fullerene derivatives, modified porphyrins and dendrimers. Such large molecules are interesting because they might serve as new modalities to address drug design problems which cannot be solved by small molecule drugs, for example blocking protein-protein interaction sites or delivering siRNA cargos into cells.[71] Unfortunately, PubChem and ChEMBL do not offer many options to explore these larger molecules. For instance, no overview of the database contents is provided, and the similarity search tools currently available on the respective websites focus on substructures, which is not well suited when relatively large molecules such as peptides are used as queries. An overview and searching across the entire content of this diverse family of large molecules is also not possible through specialized databases of biomolecules,[72] such as those for peptides,[73,74] oligonucleotides,[75] lipids,[76,77] or glycans.[78–80] Furthermore, the descriptions of chemical spaces for large molecules to date have remained focused on specific classes such as peptides and peptide macrocycles.[81,68]

Here we address this problem by designing web-based interactive tools to explore large molecules in PubChem and ChEMBL. We focus on molecules breaking at least one of the four Lipinski constraints for oral bioavailability (rule of 5: Molecular weight MW ≤ 500, number of hydrogen bond donor atoms HBD ≤ 5, number of hydrogen bond acceptor atoms HBA ≤ 10, calculated octanol/water partition coefficient clogP ≤ 5). [13] Although many orally available drugs, including peptides in particular, largely exceed Lipinski's rule of 5 limits,[82–85] Lipinski's criteria represent a useful definition to identify molecules that are clearly different from classical small molecule drugs. This concerns seven million of the 94 million entries in

PubChem and 180,185 of the nearly 2 million entries in ChEMBL 24.1, which are described herein as the non-Lipinski PubChem (NLP) and non-Lipinski ChEMBL (NLC).[83,86]

To describe NLP and NLC, we aimed to create an interactive map of the databases and a similarity search tool to identify analogs of user-defined query molecules. We have previously reported interactive 2D- and 3D-maps and similarity search tools for a variety of small molecule databases.[24,46,87–91] In these applications, composition fingerprints such as MQN (Molecular Quantum Numbers)[92] and SMIfp (SMILES fingerprint)[93] provided readily interpretable maps when projected by principal component analysis (PCA).[94] Composition fingerprints also provide interesting associations between molecules in similarity search tools.[95] However, maps and similarity searches based on these composition fingerprints are not well suited for larger molecules. For example, they do not distinguish between peptides of different sequences if they have the same amino acid composition.

To obtain a meaningful classification of NLP and NLC, we use the principle of atom-pair fingerprints, which consider pairs of atoms and the distance separating them as structural features, and assign these features to bit values either by hashing or by counting.[10,43,88,96]. Atom pair fingerprints tailored for small molecules such as CATS,[43] Xfp,[9] and 3DXfp[97] have been shown to represent molecular shape and pharmacophores. Furthermore, we have already used atom pair fingerprints successfully to describe large molecules, in one case for detailed comparisons of 3D-models of biomacromolecules such as proteins and nucleic acids in the Protein DataBank (PDB),[98] and in the second case to perform virtual screening in libraries of peptide dendrimers and bicyclic peptides.[4–6] In the latter case our atom-pair fingerprint analysis perceives meaningful differences between peptides of identical composition but different sequences.

Here we introduce a new fingerprint denoted MXFP (macromolecule extended atom-pair fingerprint), which counts atom pairs in seven different categories up to topological distances exceeding 300 bonds. We show that MXFP is well suited to describe NLP and NLC in the form of two web-based applications. First, we present interactive chemical space maps based on MXFP featuring an easily interpretable classification of NLP and NLC. Second, we report a similarity search tool identifying analogs of any query molecules based on MXFP similarity. These tools offer an unprecedented insight into the contents of NLP and NLC and reveal associations between large molecules which are otherwise difficult to identify.

2.2 Methods

2.2.1 Non-Lipinski subsets

Compound ID (CID) and SMILES were extracted for each entry in the PubChem Compound Database (downloaded April 5, 2018). For each entry, if more than one molecule was present, only the biggest fragment SMILES (based on its length) was considered for property and fingerprint calculations, however the entire SMILES was preserved. The SMILES were protonated (pH 7.4) using ChemAxon MajorMicrospecesPlugin (https://chemaxon.com). Hydrogen bond donor and acceptor count, cLogP and MW were computed for the largest fragment in each entry using RDKit, with *Lipinski*, *Descriptors* and *Crippen* modules respectively. All molecules violating more than one Lipinski rule were then classified as non-Lipinski. This led to 7,132,623 entries forming NLP and 183,185 entries forming NLC.

2.2.2 Property calculation

For each NLP and NLC entry atoms were classified into the following categories: heavy atom (HA), hydrophobic (HY), aromatic (AR), hydrogen bond acceptor and donor (HBA, HBD), positively and negatively charged (POS, NEG). AR, and HBA/HBD were assigned with, respectively, the

ChemAxon TopologyAnalyzerPlugin, and the ChemAxon HBDAPlugin. HY was assigned to aromatic carbons, halogens, sulfur atoms without heteroatom neighbors, and to carbon atoms with at least one hydrogen atom neighbor. POS and NEG were assigned based on the atom formal charge.

2.2.3 Fingerprint calculation

MXFP is a 217D atom pair topological distance fingerprint calculated using an in-house Java program in a similar manner to our previously reported atom pair fingerprints 3DP and 2DP tailored for peptides and proteins.[4–6,98] Topological distances are measured using the TopologyAnalyzerPlugin provided as part of the JChem library by ChemAxon. There are seven atom categories: heavy atom (HA), hydrophobic (HY), aromatic atoms (AR), hydrogen bond acceptor and donor (HBA, HBD), positively charged (POS) and negatively charged (NEG), and only same-category atom pairs are considered. Each of the 217 values is the sum of contributions of atom pairs at a given distance for a given atom category. For each category C, all possible atom pairs jk contribute the value $g_{jk}(d_i)/s_{jk}$ to each of the 31 distance bins value v_{Ci} as described in Equation 2.

$$g_{jk}(d_i) = e^{-\frac{1}{2}\cdot\left(\frac{d_i - d_{jk}}{d_{jk}\cdot 0.09}\right)^2}$$

$$s_{jk} = \sum_{i=0}^{30} g_{jk}(d_i)$$

$$v_{Ci} = \frac{100}{{N_C}^{1.5}} \sum_{j=1} \sum_{k=1} \frac{g_{jk}(d_i)}{s_{jk}}$$

v = MXFP bin value
C = category
$C \epsilon \{HA, HY, AR, HBA, HBD, POS, NEG\}$
$i = \{i | i \epsilon \mathbb{N} \wedge 0 \leq i \leq 30\}$
N_C = total number of atoms in category C
$d_i = \{i | 0 \leq i \leq 6\}$
$d_i = \{d_{i-1} \cdot 1.18 | 7 \leq i \leq 30\}$
d_{jk}: topological distance between atoms j and k

Equation 2

Atom pair distances d_{jk} are topological distances counted in bonds through the shortest path between two atoms. For each atom pair jk, $g_{jk}(d_i)$ is the value at distance d_i of a Gaussian of 18 % width centered on d_{jk} (Figure 4a). Gaussian values $g_{jk}(d_i)$ are sampled at the following 31 distance values d_i: 0, 1, 2, 3, 4, 5, 6, 7.1, 8.4, 9.9, 11.6, 13.7, 16.2, 19.1, 22.6, 26.6, 31.4, 37.1, 43.7, 51.6, 60.9, 71.8,

84.8, 100.0, 118.0, 139.3, 164.4, 193.9, 228.9, 270.0, 318.7. Each of these 31 gaussian values is normalized to the sum of all 31 values, s_{jk}, so that each atom pair contributes equally to the fingerprint. The sum of normalized gaussian contributions from all atom pairs of a certain atom category at distance d_i, is normalized by the number of category atoms to the power 1.5 to reduce the sensitivity of the fingerprint to molecule size, multiplied by 100 and rounded to unity to give the final fingerprint bit value v_{ci}. The 31 fingerprint bit values from each of the 7 atom categories are finally corrected by a category specific factor and joined, yielding the 217D fingerprint vector. In this work, we corrected the fingerprint bit values for the heavy atoms (HA) and aromatic atom (AR) categories by a factor 0.5 because the bit values were too high relative to the other atom categories. We calculated MXFP for the largest fragment in each NLP entry, but retained the complete SMILES in each entry.

2.2.4 Linearity calculation

The linearity of molecule *m*, *L(m)*, is a descriptor derived from MXFP. *L(m)* is defined as the ratio of *w(m)* and *w(a)*, where *a* is the linear alkane with the same number of heavy atoms as *m*, and *w* is the weighted mean of MXFP HA category, calculated according to Equation 3.

$$w = \frac{\sum_{i=0}^{30}(i+1)\cdot v_{HAi}}{\sum_{i=0}^{30} v_{HAi}}, \quad L(m) = \frac{w(m)}{w(a)}$$

w = weighted mean of MXFP HA category
$i = \{i | i \epsilon \mathbb{N} \wedge 0 \leq i \leq 30\}$
$v_{HA} = MXFP$ bin value in HA category
m = analysed molecule
a = linear alkane with m HAC

Equation 3

2.2.5 Similarity map calculation

Reference molecules were selected by sampling NLP across value triplets (HAC, AR/HAC, linearity) covering the range of each of these three descriptors in 10% value increments. One

compound was selected at random in each of the 1,000 resulting value triplets, which provided 324 reference molecules (676 of the value triplets did not contain any entry). For each database entry, the city-block distance in the MXFP chemical space, CBD_{MXFP}, to each of these 324 reference molecules was then calculated, giving a 324D NLP similarity space. The same approach was used to select reference molecules for NLC. Of the 1,000 triplets, 800 were discarded as they were not occupied by any entry, leading to a 200D NLC similarity space. The 324 NLP and 200 NLC references have very diverse structures (SMILES are provided in the SI).

2.2.6 Visualization in Faerun

The first three PCA components of the NLP and NLC similarity spaces were visualized in Faerun (variance covered, respectively: PC1 49%, PC2 28%, PC3 8%, and PC1 70%, PC2 15%, PC3 6%). A plain text file containing CID, SMILES, fingerprint, and properties of each NLP entry was processed using the Faerun preprocessing chain, which also includes a PCA service. Then Faerun was run using a docker container (https://github.com/reymond-group/faerun). Color coding of the Fearun map was enabled for HAC, HY/HAC (hydrophobic atoms fraction), AR/HAC (aromatic atom fraction), HBA/HAC (H-bond acceptor fraction), HBD/HAC (H-bond donor fraction), POS/HAC (positive charged atoms fraction), NEG/HAC (negative charged atoms), C/HAC (carbon fraction), RBC (rotatable bond count), CY/HAC (cyclic atom fraction), MW, HBA, HBD and clogP. Fraction values of atom categories are calculated from MXFP values, carbon fraction, rotatable bonds, cyclic fraction, MW, HBA, HBD and clogP are calculated with RDKit.

2.2.7 NLP-NLC comparison

20,000 entries were randomly picked from NLP or NLC and cut in two subsets of 10,000 entries each, A and B. Five series of 10,000 CBD_{MXFP} distances were then calculated as follows: a) A_{NLP} to the entire NLC, keeping the smallest non-zero value in each case; b) A_{NLC} to the entire NLC, keeping the smallest non-zero value in each case; c) A_{NLP} to B_{NLP}; d) A_{NLC} to B_{NLC}; e) A_{NLP} to A_{NLC}.

2.2.8 Similarity Search

The similarity search tool is a Python Flask (http://flask.pocoo.org/) app which uses Annoy(https://github.com/spotify/annoy) to search the MXFP NLP and NLC chemical spaces. Annoy is a C++ library with Python bindings developed by Erik Bernhardsson. Given its high speed and low memory requirements, Annoy was used to create two separate Annoy search files of NLP and NLC (for both, using *n_trees* = 50, *matrix* = Manhattan). In each similarity search instance, the user chooses to search NLP or NLC, and the correspondent Annoy file is selected. The Annoy file is used by the web app (with *search_k* = default) to retrieve the compound IDs of a pool of nearest neighbors (the no. of molecules to retrieve is a user choice). Then the compound IDs are associated back to the correspondent PubChem or ChEMBL SMILES. The results are displayed using SmilesDrawer.[99] The Similarity Search code is available open source at https://github.com/reymond-group/SimilaritySearch.

2.3 Results and Discussion

2.3.1 Non-Lipinski subsets

We define non-Lipinski molecules as those breaking at least one of the four Lipinski criteria (MW ≤ 500, HBD ≤ 5, HBA ≤ 10, clogP ≤ 5). For each PubChem and ChEMBL entry we applied the analysis to the largest molecular fragment, ignoring counter ions in the case of salts.

When applied to the currently 94 million PubChem entries, these criteria selected 7,132,623 entries, which are defined here as NLP. NLP is a diverse set, with MW spanning from 181.15 Da to 19511.8 Da, clogP from -219.4 to +132.4, HBA from 0 to 442, and HBD from 0 to 235 (Figure 3, in green). The same analysis applied to ChEMBL led to 183,185 molecules, defined here as NLC, with MW spanning from 298.1 to 10173.49 Da, clogP from -67.9 to +101.8, HBA from 0 to 286, and HBD from 0 to 124 (Figure 3, in magenta).

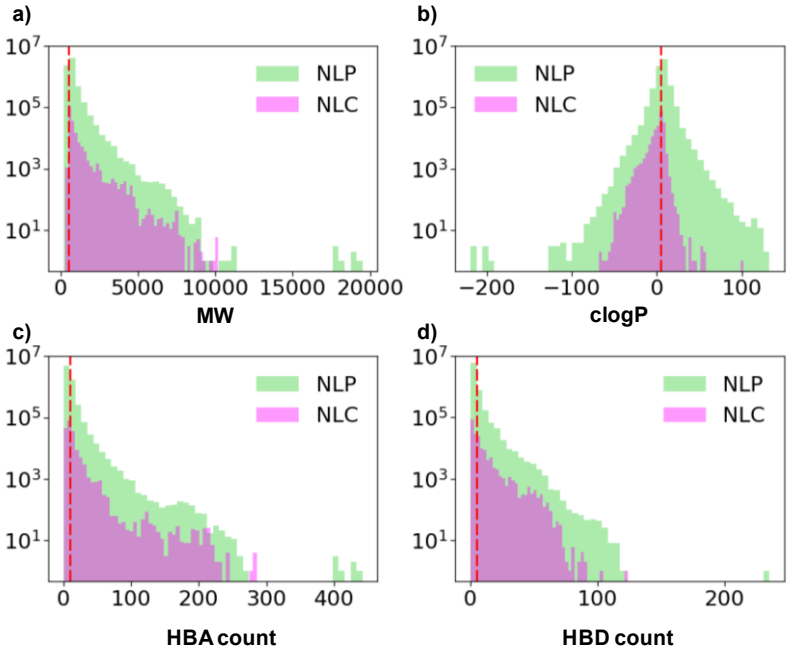


Figure 3. 1D-Histograms of NLP (green) and NLC (magenta). a) MW, b) clogP, c) HBA, d) HBD. the vertical red dashed line indicates Lipinski's rule thresholds (MW = 500 Da, clogP = 5, HBA = 10, HBD = 5).

2.3.2 Macromolecule extended atom-pair fingerprint MXFP

MXFP is a 217D fingerprint counting atom-pairs using a fuzzy approach to assign atom-pairs to distance bins as done previously in our analysis of proteins and peptides.[4–6,98] In the case of proteins, we used an atom-pair fingerprint called 3DP which considers through-space distances between atoms in experimental 3D-structures from the Protein Databank.[98] To analyze peptides, we adapted our approach to use topological distances between residues in a related fingerprint called

2DP, with all atoms in a residue positioned at the α-carbon atom.[6] For both fingerprints we consider four categories of atom pairs deemed essential for peptides, namely all heavy atoms (HA), hydrophobic (HY), positively charged (POS), and negatively charged atoms (NEG). For the MXFP presented here, we compute exact topological distances between atoms, which is suitable for any molecule. Furthermore, we use seven atom categories by additionally computing aromatic (AR), H-bond donor (HBD), and H-bond acceptor atoms (HBA), which are important to differentiate molecules such as polycyclic aromatic hydrocarbons, oligosaccharides, and oligonucleotides. As for 3DP and 2DP, we do not consider cross-category atom pairs in MXFP.

The 7,132,623 molecules in NLP correspond to 4,753,197 unique MXFP value bins. The occupancy of the MXFP bins follows a power law distribution with 75% of the bins containing only a single NLP entry (blue line, Figure 4b). A similar molecules/MXFP-bins distribution is found for NLC, where the 183,185 molecules correspond to 153,616 unique MXFP values bins (green line, Figure 4b).

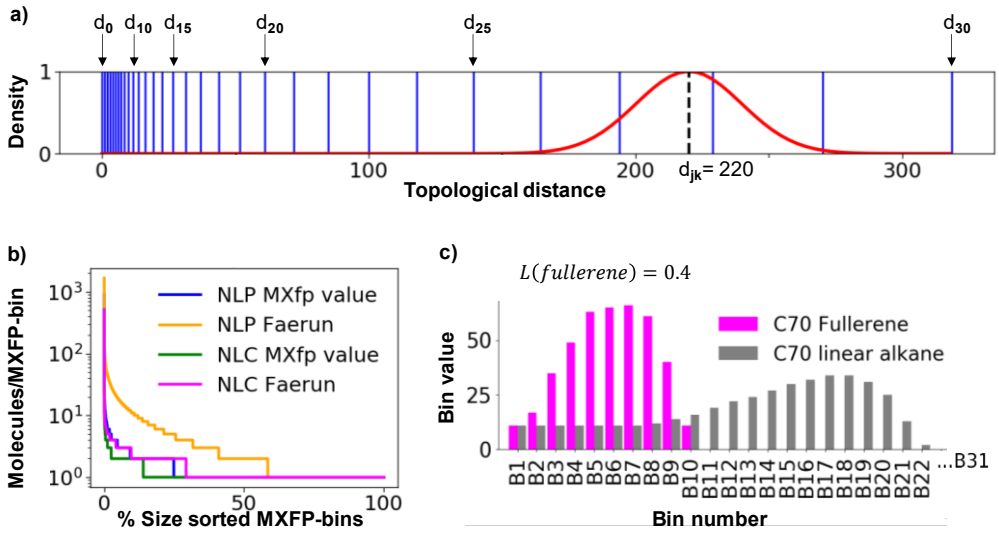


Figure 4. (a) In red Gaussian g_{jk} for an atom pair at topological distance d_{jk} = 220. In blue the 31 distances d_0 to d_{30} at which g_{jk} is sampled for calculating the contributions to MXFP. (b) Distribution of database entries in the unique MXFP-value bins (NLP in blue, NLC in green) and in the Faerun bins (NLP in orange, NLC in magenta). (c) Values of the first 31 MXFP bits for C70 Fullerene (magenta) and C70 linear alkane (grey).

The multiply occupied MXFP bins mostly contain entries sharing the same largest molecular fragment, or molecules with different structures but identical MXFP values such as diastereomeric carbohydrates (MXFP does not consider stereochemistry), or molecules with identical frameworks but different degrees of unsaturation such as lipids with fatty acids of equal length but different numbers of double bonds (MXFP does not distinguish non-aromatic carbon atoms with different degree of unsaturation). Note that grouping salts of the same compound with different counter ions, diastereoisomers of the same molecule or molecules only differing in the number of non-aromatic double bonds, makes perfect sense in the perspective of an analysis aiming at providing an overview of the database rather than a unique identifier for each entry. MXFP values are calculated using the same approach and the same parameters as used previously for 3DP and 2DP. In detail, each atom pair is converted to a Gaussian of 18 % width centered at the atom pair topological distance, which is the shortest path between the two atoms counted in bonds. This Gaussian is then sampled at 31 distances d_i spanning from $d_0 = 0$ to $d_{30} = 317.8$ bonds at exponentially increasing intervals (Figure 4a). The sampled Gaussian values are normalized and added to the MXFP distance bins for the corresponding atom-pair category, and distance bins of each category are normalized to size (Equation 2). Sampling atom-pair Gaussians at exponentially increasing distances allows to describe molecules up to a very large size using only a limited number of dimensions in the atom pair fingerprint. The approach furthermore partly erases differences between atom pairs separated by a similar number of bonds at large distances, which favors the perception of global molecular shape over structural detail.

Atom-pair fingerprints such as MXFP perceive molecular shape because spherical or cyclic molecules have a larger number of atom-pairs separated by short distances compared to linear molecules of the same size. Here we define a linearity descriptor L as a measure of topological molecular shape derived from the MXFP fingerprint. The linearity *L(m)* of

molecule *m* is defined as the ratio of the weighted mean of the heavy atom pair bin index in the MXFP of molecule *m*, *w(m)*, to the same value for a linear alkane a with the same number of heavy atoms, *w(a)* (Equation 3). The linearity value is 1 for the linear alkane, and lower for more globular molecules, e.g. *L*(fullerene) = 0.4 (Figure 4c). The linearity does not depend on building a 3D-model of the molecule as for the principal moments of inertia,[100] and is applicable to any molecule independent of its conformational flexibility.

2.3.3 MXFP chemical space visualization in Faerun

To lower the dimensionality of MXFP for visualizing NLP and NLC, we first attempted a direct principal component analysis (PCA) of the two datasets, however the first three PCs only gave partial coverage of data variance (48 % and 49%, respectively). We therefore constructed a representation based on the principle of similarity mapping.[11,101,102] Similarity mapping involves calculating similarities to a series of reference molecules in order to create a high-dimensional similarity fingerprint, which is then projected to lower dimensions by principal component analysis (PCA). The approach is interesting because the calculation of similarity maps is much faster than other dimensionality reduction methods for visualizing chemical space,[103,104] and is therefore applicable to very large datasets. Furthermore, many high-dimensional fingerprints, including MXFP, do not project well into lower dimensions if PCA is applied directly to the fingerprint values, even when adding molecules with extreme properties, called satellites, as introduced by Oprea and coworkers.[65,105] However the projection of the corresponding similarity space often produces good results.

Similarity maps calculated by randomly choosing a few hundred reference molecules usually provide an approximately constant representation which is independent of the choice of references. However, the representation can be optimized for specific purposes by selecting the reference molecules, for example series of active compounds to visualize a structure-

activity relationship study,[106] or references obtained by sampling regularly across important molecular properties to produce an ordered overview of a dataset.[107] Here we constructed similarity maps by calculating MXFP similarities to reference molecules selected across the range of MW, aromaticity and linearity covered by the NLP and NLC datasets, and then performing PCA (see methods). The procedure gave 3D-similarity maps covering 85% (NLP) and 91% (NLC) of the data variance, which was judged as sufficient to provide a good overview of the databases.

The 3D-similarity maps were then imported into Faerun, an open-source application recently reported by our group for rendering 3D-data interactively on the web.[12g] In this application each molecule is represented as a sphere, color-coded by a selected property, while its molecular structure is displayed on hover using SmilesDrawer, a compact molecular drawing program. These interactive 3D-maps enable rapid browsing through NLP and NLC to gain an overview of their contents. As for the MXFP space itself, the distribution of NLP and NLC entries into Faerun bins follows a power law (Figure 3b, NLP orange line, NLC magenta line). The high-resolution NLP map contains a total of 1,413,817 bins, corresponding to an average of five molecules per bin. Multiple occupancies per bin in the NLP similarity map occur in part for the same reasons as for the MXFP bins, but also due to rounding of coordinates in the similarity map since bins (spheres) are placed on a 500 × 500 × 500 grid. The similarity map of NLC contains 123,878 bins, with an average of 1 molecule per bin. Note that each bin (sphere) in the Faerun map can be opened in a separate tab showing the distribution of molecules in the similarity space at higher resolution.

The MXFP-similarity 3D-maps of NLP and NLC are best inspected by using the web-based view (http://faerun.gdb.tools/). We have color-coded these maps according to different descriptors from lowest (blue) to highest (magenta) value (see methods for details). A selection

of images of these color-coded representations illustrate the organization of NLP (Figure 5) and NLC (Figure 6) in the MXFP similarity space.

NLP forms a curved 3D-shape resembling a wave in which the smallest molecules are grouped on one side of the wave's head, intermediate sized molecules occupy the rest of the wave's head and the wave's body, and the largest molecules form the wave's tail, as illustrated by color-coding according to molecule size (Figure 5a). Color-coding by aromatic atom fraction shows that the outer shell of the wave's head contains molecules with the highest fraction of aromatic atoms (magenta), which are mostly polycyclic hydrocarbons (Figure 5b). The same view shows that the inner shell along the entire wave contains molecules with very few aromatic carbon atoms (blue), which comprise many linear alkanes, polyethyleneglycols, polyamines, as well as peptides. The intermediate layer contains molecules with intermediate aromatic atom fraction values (green), which are linker-extended drug-type molecules in the wave's head containing the lower size range, and oligonucleotides at the edge of the wave's tail containing the largest molecules. Oligonucleotides at the wave's tail are well visible in the map, color-coded by the fraction of negatively charged atoms (Figure 5c). This map also shows a group of smaller and more compact anionic molecules within the wave's head, which correspond to a variety of aliphatic polyphosphates and polycarboxylates.

The NLP similarity map also separates molecules according to their shape as measured by the MXFP derived linearity descriptor *L* discussed above (Figure 5d). The narrow blue region at the wave's head corresponds to globular molecules with a high percentage of aromatic carbons such as fullerenes. A second narrowly defined region at the center of the inner shell is colored in magenta and features strictly linear molecules containing long alkyl or polyethylene-glycol chains without any branching points. Peptides and oligonucleotides appear at intermediate values of linearity (yellow), which reflects the fact that these molecules are

multiply branched by the attachment of amino acid side-chains (peptides) and nucleosides (oligonucleotides) along the main peptide respectively phosphodiester chain.

The different compound families are nicely separated by color-coding by the fraction of carbon atoms (Figure 5e). The figure shows examples of polycyclic hydrocarbons (exabenzocoronene, carbon fraction = 1.0, magenta), carbohydrates (difucosyllacto-N-hexaose, carbon fraction = 0.56, blue), peptides (exenatide, carbon fraction = 0.62, green) and oligonucleotides (mipomersen sodium, carbon fraction = 0.50, blue). Note that a close inspection of the MXFP similarity map of NLP using Fearun reveals many entries that are obvious mistakes in the PubChem database. For example, most mipomersen structures in PubChem are not drawn as the correct phosphorothioates but as the incorrect phosphate thioesters. Further structures of doubtful identity are also visible that contain linear chains of nitrogen and oxygen atoms.

NLC forms a similar but more sparsely populated wave-shaped 3D-similarity map. As with NLP, molecular size increases when navigating the map from the wave's head to its tail (Figure 6a, HAC color code). Aromaticity is higher in the outer shell of the wave head and diminishes upon traversing the map towards its inner shell (Figure 6b, AR/AHC color code). Compared to NLP (Figure 5b) the highly aromatic outer shell is less populated. Browsing this area in Fearun reveals an almost total absence of polycyclic hydrocarbons. As for NLP, the inner shell of the wave-shaped map contains mostly polyethylene-glycols, polyamines, and peptides, however the linear alkanes seen in NLP are mostly missing. As in NLP, the intermediate shell at the edge of the wave's tail with intermediate aromatic fraction (in green) contains oligonucleotides. Oligonucleotides are also well visible in blue using the NEG/HAC color code (Figure 6c). In terms of molecular shape, color coding by linearity *L* shows a similar distribution as for NLP (Figure 6d).

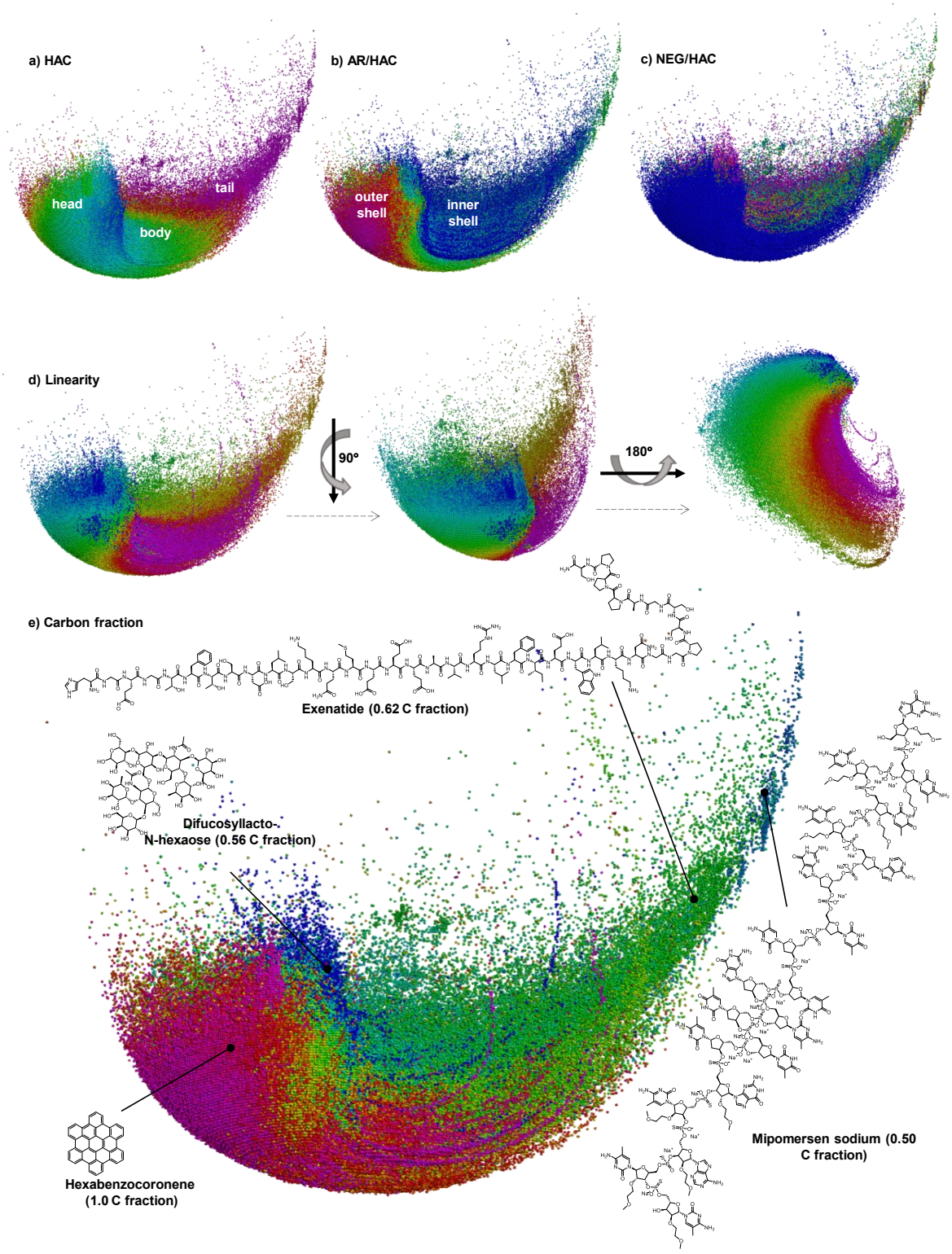


Figure 5. Similarity map of the NLP chemical Space colored using HAC (*a*), AR/HAC (*b*), NCHRG/HAC (*c*), linearity (*d*), and carbon fraction (*e*). In *d* are shown different rotation of the map. In *e* is shown the placement and the structure of hexabenzocoronene, difucosyllacto-N-hexaose, exenatide, and mipomersen sodium, as representative compounds of, respectively, polycyclic hydrocarbons, carbohydrates, peptides, and oligonucleotides.

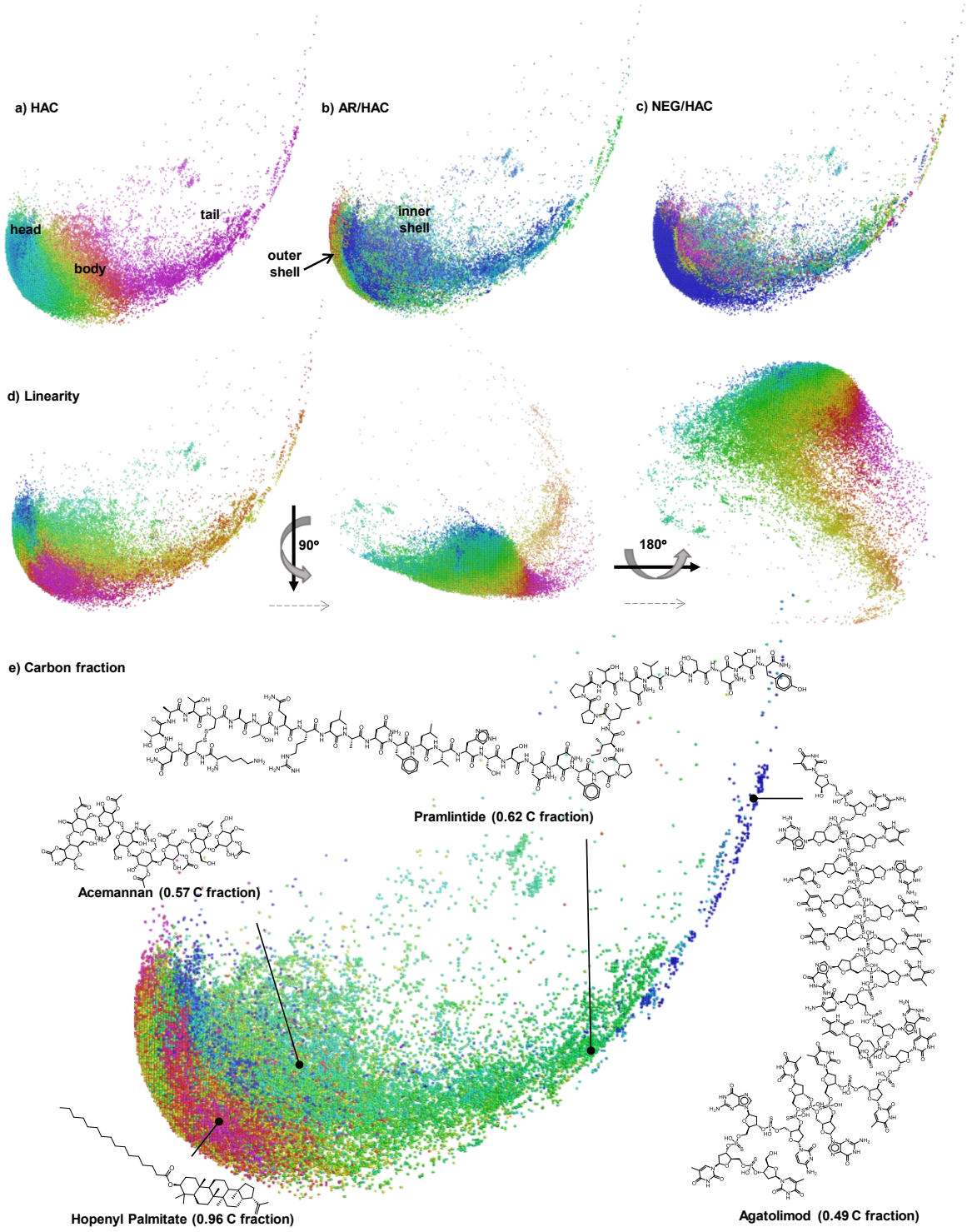


Figure 6. Similarity map on the MXFP NLC chemical Space colored using HAC (*a*), AR/HAC (*b*), NCHRG/HAC (*c*), linearity (*d*), and carbon fraction (*e*). In *d* are shown different rotation of the map. In *e* is shown the placement and the structure of hopenyl Palmitate, acemannan, pramlintide, andagatolimod, as representative compounds of, respectively, high carbon fraction molecules, carbohydrates, peptides, and oligonucleotides.

As with NLP, coloring the NLC similarity map by carbon fraction separates different compound families (Figure 6e). The figure shows examples of steroids (hopenyl palmitate, carbon fraction = 0.96, magenta), carbohydrates (acemannan, carbon fraction = 0.57, blue), peptides (pramlintide, carbon fraction = 0.62, green) and oligonucleotides (agatolimod, carbon fraction = 0.49, blue).

2.3.4 Comparing NLC with NLP

Because ChEMBL is one of the sources that feeds into PubChem, NLC represents a small (2.7%) subset of NLP.[108,109] To investigate if the remaining 97,8% of NLP cover a broader or different chemical space compared to this small NLC subset, we analyzed the CBD_{MXFP} distance distribution between 10,000 randomly picked NLP molecules and their NLC nearest neighbors, between NLC nearest neighbors, and between random pairs in NLP, NLC, and between NLP-NLC cross-pairs (Figure 7).

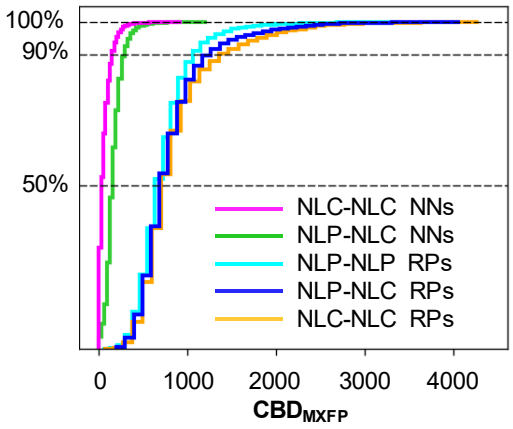


Figure 7. Distribution of CBD_{MXFP} distances between NLC and NLP molecules for nearest neighbors (NNs) and random pairs (RPs). See text and methods for details.

The analysis shows that 50% of NLP molecules have a nearest neighbor in NLC within CBD_{MXFP} 170, and more than 90% within CBD_{MXFP} 300 (Figure 7, green line), which is only

a slightly larger distance distribution compared to the distance separating NLC nearest neighbors (Figure 7, magenta line). These nearest neighbor distances are much shorter than distances between random pairs of molecules within NLP (Figure 7, cyan line), within NLC (Figure 7, orange line), or between NLP and NLC molecules (Figure 7, blue line). We conclude that NLC and NLP cover a similarly broad chemical space, that NLC represents an almost random subset of NLP, and that NLP, although being 37-fold larger than NLC, does not cover a significantly different chemical space.

2.3.5 MXFP similarity search

PubChem currently offers a similarity search window in its beta version, which provides meaningful analogs of most query molecules. Unfortunately, this search option is designed for small molecules and fails to return any analog or does not return meaningful analogs when challenged with large molecules, most often when the query molecule is not itself present in PubChem. The same issue is experienced using the search function in ChEMBL. Examples of failed searches are shown in Table 21.

Here we designed an MXFP-similarity search tool for NLP and NLC as a web-portal using the approximate nearest neighbor search Annoy (Approximate Nearest Neighbors Oh Yeah, https://github.com/spotify/annoy) (Figure 8). This search option allows the user to browse NLP or its subset NLC and returns hundreds of MXFP-analogs of a query molecule in approximately 30 second per query.

The MXFP similarity search often returns results comparable to those provided by the PubChem and ChEMBL webpages whenever matched molecules have comparable substructures. However, compared to the PubChem and ChEMBL websites, which often fail to return results for unusual queries, MXFP similarity search provides a list of analogs in all cases. Analogs identified by MXFP similarity often comprise molecules with an overall

molecular shape comparable to the query molecules, but with different structural composition. A good example is provided by searching NLP and NLC for analogs of **T7**, an antimicrobial peptide dendrimer with an unusual multi-branched peptide architecture which is active against multidrug resistant Gram-negative bacteria.[4] While the PubChem and ChEMBL webpages do not find any meaningful analogs for this query, returning smaller and linear peptides, our MXFP similarity search points to related polycationic dendritic molecules of very different detailed structure. Besides many structures coming from patents, one example of interest at rank 4 in the search is CID 49775868, which is a peptide derivatized dendrimer of overall similar size, charge and shape as **T7**, but with very different detailed structure, reported to be active against HIV (Figure 8).[110]

2.4 Conclusion

Here we focused on developing interactive tools to visualize and search large molecules in PubChem and ChEMBL breaking at least one of Lipinski's constraints for bioavailbility, defined here as NLP (7 million molecules) and NLC (180 185 molecules). We defined a 217D atom-pair fingerprint, MXFP, to describe these molecules in terms of molecular shape and pharmacophores. While MXFP is in principle suitable to describe molecules across the entire size range, here we focused on using this fingerprint to represent NLP and NLC in an interactive 3D-map and to enable a similarity search tool. These tools allow to rapidly browse through these diverse collections of macromolecules with unprecedented efficiency and identify interesting compound families and similarities between molecules which are otherwise difficult to perceive.

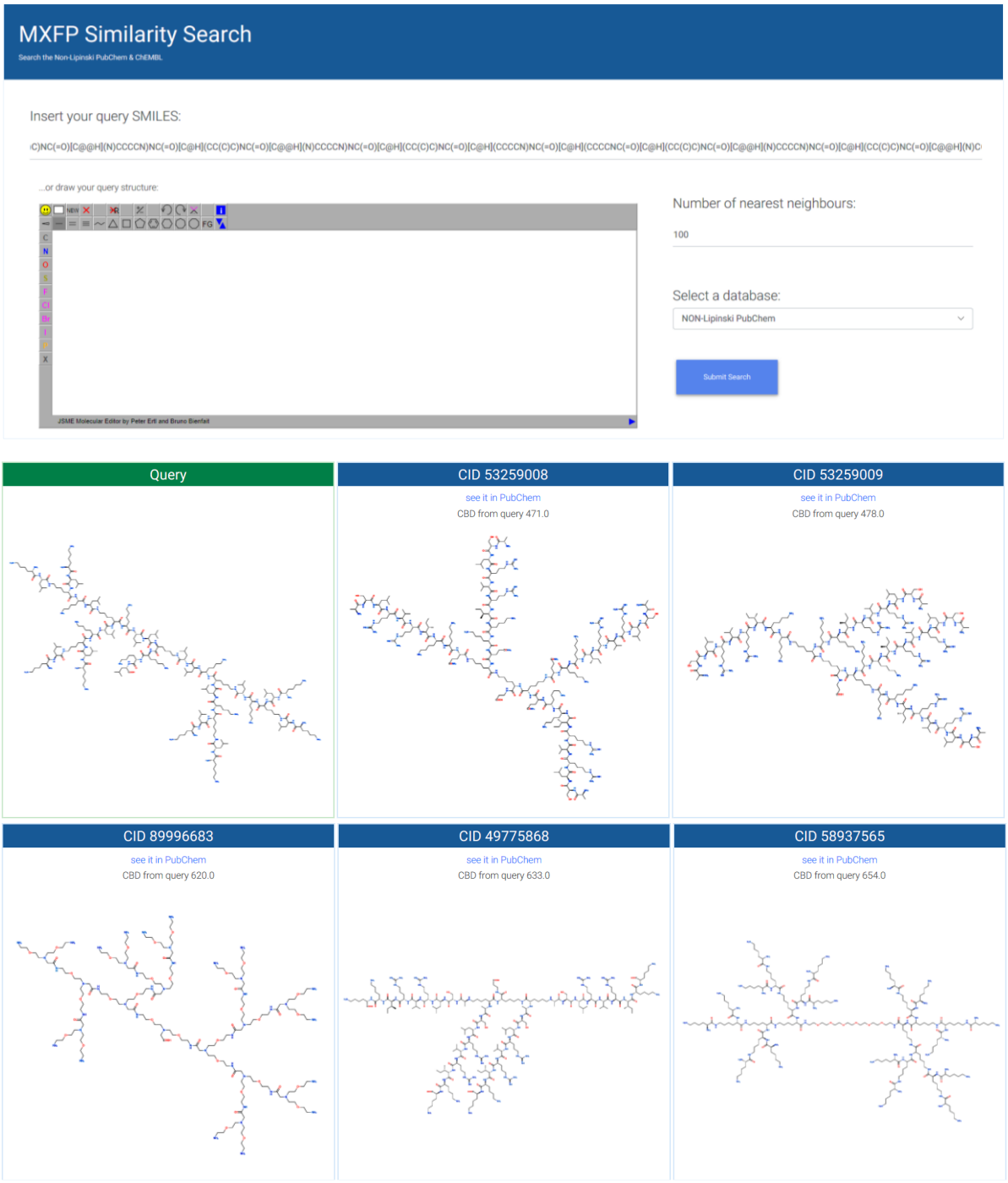


Figure 8. MXFP web interface. MXFP similarity search results for peptide dendrimer **T7**.

Chapter Three – One Molecular Fingerprint to Rule them All: Drugs, Biomolecules, and the Metabolome

This work is based on the peer-reviewed publication:

Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. J. Cheminformatics 2020, 12 (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

Abstract

Background: Molecular fingerprints are essential cheminformatics tools for virtual screening and mapping chemical space. Among the different types of fingerprints, substructure fingerprints perform best for small molecules such as drugs, while atom-pair fingerprints are preferable for large molecules such as peptides. However, no available fingerprint achieves good performance on both classes of molecules.

Results: Here we set out to design a new fingerprint suitable for both small and large molecules by combining substructure and atom-pair concepts. Our quest resulted in a new fingerprint called MinHashed atom-pair fingerprint up to a diameter of four bonds (MAP4). In this

fingerprint the circular substructures with radii of $r = 1$ and $r = 2$ bonds around each atom in an atom-pair are written as two pairs of SMILES, each pair being combined with the topological distance separating the two central atoms. These so-called atom-pair molecular shingles are hashed, and the resulting set of hashes is MinHashed to form the MAP4 fingerprint. MAP4 significantly outperforms all other fingerprints on an extended benchmark that combines the Riniker and Landrum small molecule benchmark with a peptide benchmark recovering BLAST analogs from either scrambled or point mutation analogs. MAP4 furthermore produces well-organized chemical space tree-maps (TMAPs) for databases as diverse as DrugBank, ChEMBL, SwissProt and the Human Metabolome Database (HMBD), and differentiates between all metabolites in HMBD, over 70 % of which are indistinguishable from their nearest neighbor using substructure fingerprints.

Conclusion: MAP4 is a new molecular fingerprint suitable for drugs, biomolecules, and the metabolome and can be adopted as a universal fingerprint to describe and search chemical space. The source code is available at https://github.com/reymond-group/map4 and interactive MAP4 similarity search tools and TMAPs for various databases are accessible at http://map-search.gdb.tools/ and http://tm.gdb.tools/map4/.

3.1 Introduction

The diversity and size of the organic molecules of possible interest as drugs steadily increases as medicinal chemistry addresses ever more complex biological processes while also exploiting the expanding scope of synthetic organic chemistry.[111–113] Cheminformatics enables the exploitation and understanding of this diversity by describing molecules as molecular fingerprints, encoding their structural characteristics as a vector.[114,115] These fingerprints can be used for fast similarity comparisons forming the basis for structure-activity relationship studies, virtual screening, and the construction of chemical space maps.[45,116–118]

Most molecular fingerprints have been conceived, validated, and used in the context of small molecule drugs within the classical Lipinski limits, [13] and are not well suited to describe larger molecules. For instance, the most popular molecular fingerprint is the Morgan fingerprint, also known as extended-connectivity fingerprint ECFP4.[7] ECFP4 belongs to the best performing fingerprints in small molecule virtual screening[16] and target prediction benchmarks,[119,120] together with the related MinHashed fingerprint MHFP6.[8] Both fingerprints perceive the presence of specific circular substructures around each atom in a molecule, which are predictive of the biological activities of small organic molecules. However, both have a poor perception of the global features of molecules such as size and shape. They also fail at perceiving structural differences that may be important in larger molecules, such as distinguishing between regioisomers in extended ring systems (e.g. 2,7- versus 2,8-dichlorodioxin), between linkers of different lengths, or between scrambled peptide sequences of identical composition and length.

The above limitations can be addressed by using atom-pair fingerprints, [10] which encode molecular shape and are often used for scaffold-hopping.[121,97,9] We have shown that atom-pair fingerprints are suitable to describe large molecules by mapping the Protein

DataBank.[98] We also used atom-pair fingerprints to discover and optimize novel antimicrobial peptides in virtual libraries of bicyclic peptides[5,6] and peptide dendrimers,[4,122] to create chemical space maps[24] of molecules beyond the Lipinski limit found in the PubChem and ChEMBL databases,[14] and to drive a genetic algorithm to produce analogs of peptides with diverse chain topologies.[25] Overall, atom-pair fingerprints have an excellent perception of molecular shape for both large and small molecules and overcome the above-mentioned limitations. However, they do not encode molecular structure in detail and perform poorly in small molecule benchmarking studies compared to substructure fingerprints such as ECFP4 and MHFP6.

Here we set out to investigate if the atom-pair approach could be combined with circular substructures as implemented in the above mentioned MinHashed fingerprint MHFP6 to create a new fingerprint suitable for small molecule virtual screening but also capable of describing large molecules including biopolymers such as peptides. Such a fingerprint would provide an elegant unified description of molecules across very different sizes and might also be useful to describe molecules of intermediate size such as large natural products and metabolites. Our quest uncovered a new fingerprint which we call MAP4 (MinHashed Atom-Pair fingerprint up to four bonds). MAP4 encodes atom pairs and their bond distance similarly to the AP fingerprint implemented by RDKit,[123] however in MAP4 atom characteristics are replaced by the circular substructure around each atom of the pair, written in SMILES format. MAP4 uses the same MinHashing technique as MHFP6, a principle borrowed from natural language processing which enables fast similarity searches in very large databases by locality sensitive hashing (LSH). LSH is a technique that allows the creation of self-tuning indexes, which are then used to generate a forest of trees that can be traversed for an approximate but fast similarity search.[124,125,42,126]

We show that MAP4 outperforms substructure fingerprints in small molecule benchmarking studies[16] and at the same time outperforms other atom-pair fingerprints in a peptide benchmark designed to evaluate performance on large molecules. Furthermore, we show with the example of various interactive tree-maps (TMAPs) [12] that MAP4 has excellent properties to map the chemical space of databases of molecules of interest across the life sciences such as bioactive molecules of various sizes (DrugBank,[127] ChEMBL,[128] non-Lipinski ChEMBL),[14] peptides (peptides up to 50 residues from SwissProt),[129,130] and metabolites (Human Metabolome database).[17]

3.2 Methods

3.2.1 Fingerprint calculation

The MinHashed Atom Pair (MAP) fingerprint calculation requires a canonical and anisomeric SMILES representation of the input molecule, as well as the parameter *r*, which signifies the maximal radius of the circular substructures to be considered (default radius value $r = 2$ corresponding to a diameter $d = 4$ for MAP4). The fingerprint is calculated as follows: First, the circular substructures surrounding each non-hydrogen atom j in the molecule at radii 1 to r are written as canonical, non-isomeric, and rooted SMILES string $CS_r(j)$ using RDKit.[131] Second, the minimum topological distance $TP_{j,k}$ separating each atom pair (j, k) in the input molecule is calculated. Third, all atom-pair shingles $CS_r(j)|TP_{j,k}|CS_r(k)$ are written for each atom pair (j, k) and each value of r, placing the two SMILES strings $CS_r(j)$ and $CS_r(k)$ in lexicographical order (Figure 9). Fourth, the resulting set of atom-pair shingles is hashed to a set of integers S_i using the unique mapping SHA-1,[40] and its corresponding transposed vector S_i^T is finally MinHashed to form the MAP4 vector Equation 4). A detailed description of the MinHash method used here can be found in our recent publication on MHFP6.[8]

$$\text{col\_min} \rightarrow \text{returns the smallest number in each column}$$
$$a, b \rightarrow \text{randomly generated vectors of same length}$$
$$a_i, b_i \epsilon \{0, \ldots 2^{32} - 1\}$$
$$m = 2^{32} - 1 \text{ (maximum Hash)} \qquad \text{Equation 4}$$
$$p = 2^{61} - 1 \quad \text{(Mersenne prime)}$$
$$\text{hmin}(s_i, a, b) = \text{col\_min}\left(\left((a \times s_i^T + b) \bmod p\right) \bmod m\right)$$

In this work, we investigate twelve different variations of the atom pair MinHashed fingerprint considering different shingle radii *r* as MAP2 (*r* = 1), MAP4 (*r* = 2), MAP6 (*r* = 3), and MAP8 (*r* = 4), each of them in a 1024-dimensions and 2048-dimensions versions, as well as 2048-dimensions folded (instead of MinHashed) variants using the modulo operation in form of foldedAP2 (*r* = 1), foldedAP4 (*r* = 2), foldedAP6 (*r* = 3), and foldedAP8 (*r* = 4).

3.2.2 Peptide benchmark datasets

Thirty random linear sequences (ten 10-mers, ten 20-mers, and ten 30-mers) were generated with each of all 20 proteogenic amino acids picked with the same probability (Table 22). For each sequence, we produced 10,000 scrambled unique versions using all amino acids of the parent sequence in random different combination. We also produced 10,000 mutated unique versions by considering the sequence length as the maximum number of possible mutated residues, and for each possible number of point mutations, we generated *n* mutated sequences, where *n=ceiling(10,000/maximum number of possible mutations)*; if more than 10,000 sequences were produced, only the first 10,000 were selected. The scrambled and the mutated sets were searched with BLAST[124] using the original sequence as a query. The search was performed with blastp using default settings (Gap opening penalty = 11, Gap extension penalty = 1, Expectation value = 10.0, Word size = 3, Max scores = 25, Max alignments = 15, Query filter = SEG, Matrix = blosum62). The resulting BLAST analogs (Expectation value < 10.0) were labelled as active, while the remaining sequences were labelled as decoys. The protonated

SMILES of all peptide sequences were generated using a method of the recently published Peptide Design Genetic Algorithm (PDGA). [25] To generate the extended fingerprint benchmark training lists for each peptide dataset, 50 different sets of 5 actives and 10% of decoys were randomly picked and stored using the Python package pickle. The peptide active and inactive datasets and the training lists can be found at https://github.com/reymond-group/map4.

3.2.3 Benchmark metrics and parameters

To evaluate the fingerprints in the extended benchmark, we used the following metrics: AUC, EF1, EF5, BEDROC20, BEDROC100, RIE100, and RIE20. The virtual screening was repeated five times with five different queries. To assess similarity (or dissimilarity) among molecules in the benchmark virtual screenings, we used the Jaccard similarity for MinHash-based fingerprints, Manhattan distance for the 217-dimensions atom-pair fingerprint MXFP (macromolecule extended atom-pair fingerprint), and Dice similarity in all other cases. Details regarding the benchmark implementation can be found in the 2013 Riniker et. al. publication.[16]

3.2.4 Similarity Search databases preprocessing

ChEMBL 25.0 and Metabolome 4.0 were extracted and manipulated as follows: (1) All structures were canonicalized and chirality information was removed using RDKit; (2) fragments were removed; (3) Heavy atoms were counted using RDKit and compounds with less than 2 heavy atoms were discarded. The filtering resulted in 1,699,888 and 96,456 unique SMILES for the ChEMBL and Metabolome datasets respectively. For ChEMBL molecules, activity information was extracted if present but only when the confidence score was above 5 for a standardized value ≤ 10,000 nM. In the Human Metabolome database preprocessing, the metabolite source was always annotated if available. Natural peptide sequences with 50 of fewer residues were extracted from the SwissProt dataset and translated into non-chiral SMILES using PDGA, [25] resulting in 9,054 unique structures.

The three datasets were encoded with MAP4 and MHFP6 in 512-dimensions. For each database and fingerprint variant, an LSH forest of 32 trees was generated using the TMAP class. These LSH forests were used as an index for the similarity search. For details on MHFP6, and LSH forest implementation please refer to the recent Probst and Reymond publications.[8,12]

3.2.5 Similarity Search Implementation

A fast similarity search tool was implemented for ChEMBL, SwissProt, and the Metabolome databases. The given query is canonicalized and chirality information is removed with RDKit. Then, the nearest neighbors of the processed query are retrieved using the LSH forest corresponding to the chosen database to search in. The query molecule can be provided as a SMILES (drawn structure or pasted SMILES in the JSME editor)[132] or as a linear sequence of a natural peptide. In the latter case, the sequence is transformed into its corresponding SMILES using PDGA as for the SwissProt database and the benchmark compounds. The code of the similarity search is available at https://github.com/reymond-group/map4.

3.2.6 Databases preprocessing for TMAP

For SwissProt, the previously mentioned similarity search LSH forest was used. ChEMBL 25.0, Metabolome 4.0, and Drugbank 5.4 were extracted and compounds with less than 2 atoms were discarded, resulting in 1,870,343, 114,016, and 10,607 SMILES for the ChEMBL, Metabolome, and Drugbank datasets respectively. A subset of the ChEMBL database was generated by random sampling of 187,034 compounds (10%). Activity information of ChEMBL molecules and sources of metabolome molecules were extracted as previously described for the Similarity Search databases. To provide a TMAP focused on the larger structures in the database, ChEMBL molecules that broke more than one Lipinski's rules of five [13] were collected to form an additional dataset containing 229,067 entries (Lipinski descriptors were calculated using RDKit).

For the SwissProt database, positive and negative charges were calculated directly from the peptide sequences: R and K counted as a positive charge each, D and E counted as a negative charge each, all other residues were considered neutral. The number of aromatic atoms (AR) was calculated counting all lowercase “c”, “n”, “s”, and “o” not belonging to a two-letter element in the canonical SMILES. All other properties were calculated using RDKit.

The five datasets were encoded with MAP4 in 512-dimensions. For each database and fingerprint version, an LSH forest of 32 trees was generated using the TMAP class. The obtained LSH forests were used to layout the corresponding TMAPs. The color-codes of property values on each TMAP (accessible via the TMAP menu) were obtained by first ranking molecules using SciPy,[133] and then assigning the rank to a color linearly along the color scale. For the property “Phosphorus count” we used a *dense* ranking, in which molecules with the same number of P atoms receive the same rank. For all other properties a standard (or *average*) ranking was used: the average of the ranks that would have been assigned to all the tied values was assigned to each value. For details on TMAP please refer to the related publication. [12]

3.2.7 Nearest neighbor analysis

The Human Metabolome data set was sorted unique after removing stereochemistry information and for each molecule, the distance from its nearest neighbor was calculated in the MAP4-1024, MHFP6-1024, TT (not hashed), AP (not hashed), and ECFP4-1024 chemical spaces. AP, TT, and ECFP4 were calculated with RDKit. In each fingerprint space, for each structure, a similarity search against the entire dataset was performed and the NN retrieved. The similarity was assessed as Tanimoto Distance calculated with RDKit.

3.3 Results and Discussion

3.3.1 Fingerprint Design

Our atom-pair fingerprint is designed similarly to the AP fingerprint implemented by RDkit. AP encodes atom pairs using atomic invariants combined with their bond distances. Instead of using atomic invariants, we use the circular environment of each atom in the pair up to a preset radius, written as canonical SMILES, similar to the method used for MHFP6. Recording circular substructures is expected to lead to a more detailed perception of substructures in the fingerprint enabling better performance in small molecule benchmarks, while the bond distance information should translate into a perception of molecular size and shape. For each radius value r (typically $r = 1$ and 2), we encode each atom pair as a character string consisting of the two canonical SMILES of the circular substructure around each atom up to the set radius and the bond distance information. We then hash these atom-pair strings and use MinHash to produce the actual fingerprint to capitalize on the advantages of this approach over binary encoding as previously demonstrated with MHFP6 (see methods, Equation 4).[8] For example, our MinHashed Atom Pair fingerprint with $r = 2$ (MAP4) encodes pairs of circular substructures with radius $r = 1$ and 2 (Figure 9).

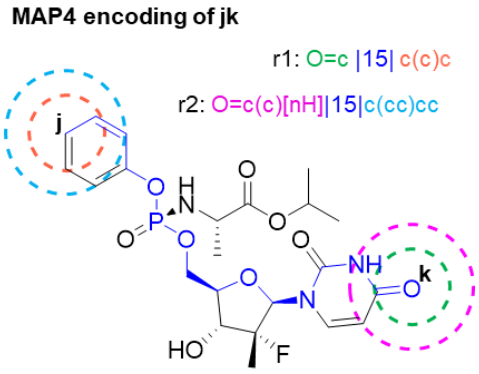


Figure 9. MAP4 atom pair encoding. The circular substructures around atoms j and k at radius $r = 1$ and $r = 2$ are written as SMILES placed in lexicographical order separated by the bond distance between the two atoms along the shortest path (blue). These character strings are the atom-pair molecular shingles for this atom-pair for $r = 1$ and $r = 2$.

3.3.2 Benchmarking study design

To evaluate the performance of MAP4 we use a modified version of the fingerprint benchmark developed by Riniker and Landrum. [16] The benchmark provides a detailed insight about the performance of an evaluated fingerprint in the recovery of actives in a virtual screening of a database of known actives and decoys, where the actives/decoys sets are taken from the DUD,[44] the MUV,[134] and the ChEMBL[128] datasets. However, since most molecules are within the rules of five limits (Figure 38), the benchmark gives no explicit information on the performance of an evaluated fingerprint in encoding larger molecules. We have therefore extended the benchmark with a series of peptides as exemplary large biomolecules not only because they are an important class of drugs, but also because their similarity can be assessed with BLAST, a reliable and widely used tool. Our peptide benchmark consists of 60 scrambled and mutated peptide datasets generated from 30 randomly generated sequences. In each set the actives and decoys are defined through their sequence similarity to the corresponded query: the BLAST analogs are labelled as active, while the remaining sequences are labelled as inactive (see methods and Table 1).

Table 1. Average number and percentage of actives in all datasets used for the benchmark.

	MUV[a)]	DUD [a)]	ChEMBL [a)]	Mutated Peptides [b)]	Scrambled Peptides [b)]
Average n.o. actives	30.0±0.0	91.3±80.5	100.0±0.0	500.2±0.7	56.0±27.4
Average % actives	0.2±0.0%	2.2±0.4%	1.0±0.0%	5.3±0.0%	0.6±0.2%

a) Known actives used in the Riniker and Landrum [16] benchmark. b) BLAST analogs of a defined query generated for this study.

We include twenty-one different fingerprints in the comparison, comprising the twelve variations of our MAP4 fingerprint as described in the methods, and nine reference fingerprints

performing particularly well for small or large molecules. This reference set includes ECFP4 and MHFP6 in their 1024-dimensions and 2048-dimensions versions as best performing fingerprints for small molecules, MXFP (macromolecule extended atom-pair fingerprint, 217-dimensions atom-pair fingerprint) as a good performing fingerprint for large molecules and peptides,[25] and the Atom Pair (AP) and Topological Torsion (TT) fingerprints from RDKit. In the AP and TT fingerprints atoms are represented using their atom type, their number of heavy neighbors, and their number of pi electrons. AP encodes all atom pairs and their distance as a number, while TT encodes all atoms along the path between two atoms up to topological distance of four bonds. Note that AP and TT are not hashed as in the original benchmark. Finally, our reference set includes MACCS and ECFP0 as baseline fingerprints following the Riniker benchmark. [16]

We use five different metrics in the benchmark, namely AUC (Figure 10a), RIE100 (Figure 39a) and RIE20 (Figure 39b), BEDROC100 (Figure 10b) and BEDROC20 (Figure 39c), and EF1 (Figure 39d) and EF5 (Figure 10c). The relative performance of the different fingerprints is then assessed by computing their average rank in each of the metrics following the Riniker approach (Figure 11a-c). The statistical relevance of the ranks is assessed with the Friedman Test provided in the Riniker benchmark, where the post hoc analysis is performed using Wilcoxon-Nemenyi-McDonald-Thompson test (Figure 40 to B.5).[135,136]

3.3.3 Benchmarking results

We first compare MAP4 with the nine reference fingerprints presented above. In the small molecule benchmark MAP4 is slightly better than substructure fingerprints (ECFP4, MHFP6, and TT), yet the difference is not statistically significant. However, MAP4 outperforms atom-pair fingerprints such as AP and MXFP, which perform significantly worse in this benchmark (Figure 11a and Figure 40). The situation is reversed in the peptide benchmark, where atom-

pair fingerprints significantly outperform substructure fingerprints (Figure 11b). MAP4 performs best among these atom-pair fingerprints, however, the difference is not statistically significant (Figure 41). Remarkably, MAP4 is the only fingerprint maintaining good performances in both benchmarks.

Having established that MAP4 outperforms other known fingerprints in the combined small molecules and peptides tasks, we next investigate if further improvements might be possible in 12 variations of the MAP4 fingerprint considering different shingle radii (r = 1, 2, 3, 4), compression methods (MinHash versus folding), and the number of dimensions (1024 or 2048). We include MHFP6-2048 and the RDKit AP as reference fingerprints in this comparison. Comparing the average fingerprint rank for small molecules (Figure 12a) and peptides (Figure 12b), as well as the performance metrics on each dataset (Figure 42) shows that the MinHashed fingerprints (MAPs) rank better than their folded versions (foldedAPs) in a statistically significant manner, except for foldedAP2 when using only the small molecule datasets (Figure 40 and B.4). The better performance of MinHashed over folded versions of the same fingerprint was already observed in our study of MHFP6,[8] and probably results from the fact that MinHashing creates fewer unintended bit collisions as compared to modulo-based hashing (folding) as an information compression method. Bit collision is most likely also the reason for the decreasing performance of foldedAPs when the radius, and therefore the encoded information, increases.

Among the different MAPs, those with larger radii perform better, however, the difference is not statistically significant. At the same time increasing the radius from r = 1 (MAP2) to r = 2 (MAP4), r = 3 (MAP6) and r = 4 (MAP8) defines an exponentially increasing number of unique atom-paired molecular shingles, as exemplified for the case of the ChEMBL database (Table 2). The selected MAP4 (r = 2) represents a compromise to represent substructures in reasonable but not exaggerated detail. In the MAP4 ChEMBL space, there are

46,430,912 atom-pair molecular shingles. While half of them are seen only once, the most common Shingle is present in 85% of ChEMBL structures (Figure 12d). Note that the radius can be selected by the user in the current implementation.

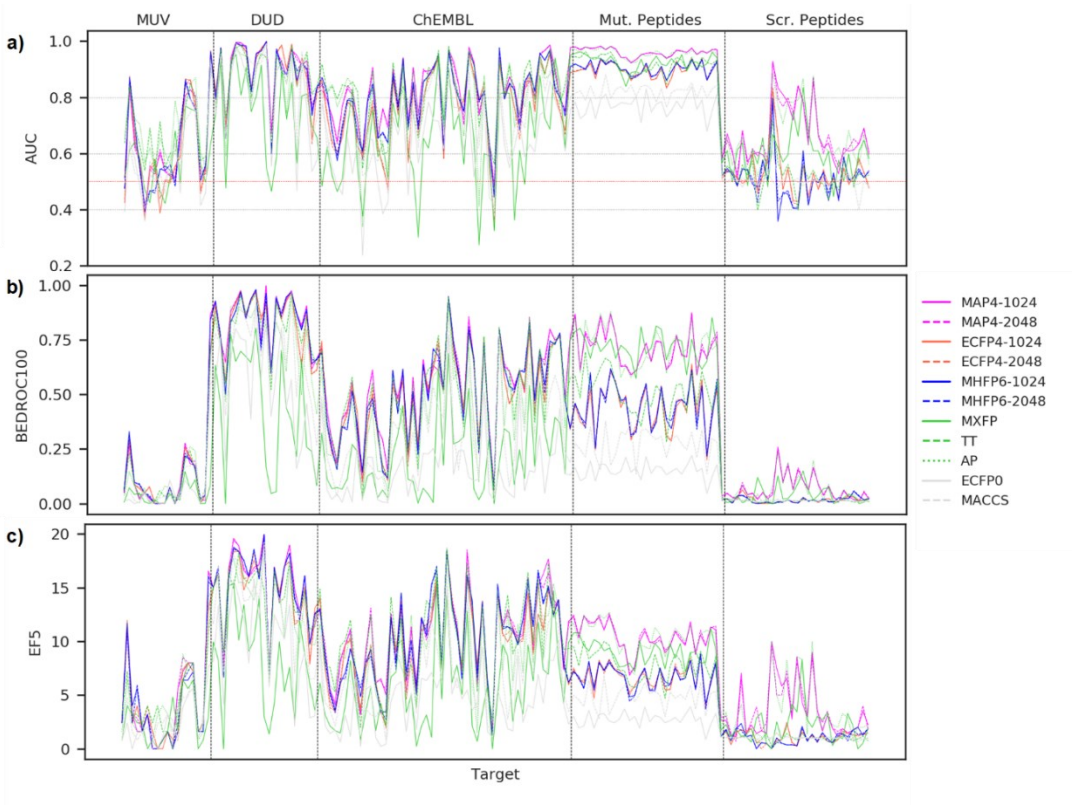


Figure 10. AUC (a), BEDROC100 (b), and EF5 (c) of MAP4 (magenta), ECFP4 (orange), MHFP6 (blue), MXFP (solid green line), TT (dashed green line), AP (dotted green line), MACCS (solid gray line), and ECFP0 (dashed gray line) across all small molecules and peptide targets (17 MUV targets, 21 DUD targets, 50 ChEMBL targets, 30 mutated peptide targets, and 30 scrambled peptide targets).

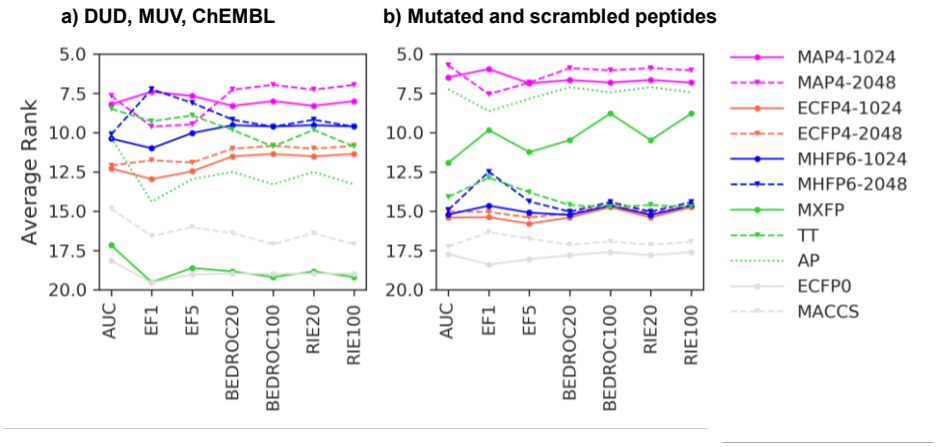


Figure 11. Average ranking of MAP4 (magenta), ECFP4 (orange), MHFP6 (blue), MXFP (solid green line), TT (dashed green line), AP (dotted green line), MACCS (solid gray line), and ECFP0 (dashed gray line) in in the fingerprint benchmark when using only small molecules datasets (17 MUV targets, 21 DUD targets, and 50 ChEMBL targets, a) and only peptide datasets (30 mutated peptide targets and 30 scrambled peptide targets, b). Note that 11 out of 21 fingerprints are shown.

Table 2. Analysis of ChEMBL using MinHashed atom-pair fingerprint variants.

Fingerprint[a)]	**Unique Shingles**[b)]
MAP2 ($r = 1$)	1,913,607
MAP4 ($r = 2$)	46,430,912
MAP6 ($r = 3$)	205,576,613
MAP8 ($r = 4$)	465,393,948

a) MinHashed atom-pair fingerprint version with different shingle radii. b) Number of different atom-paired molecular shingles in the entire ChEMBL database.

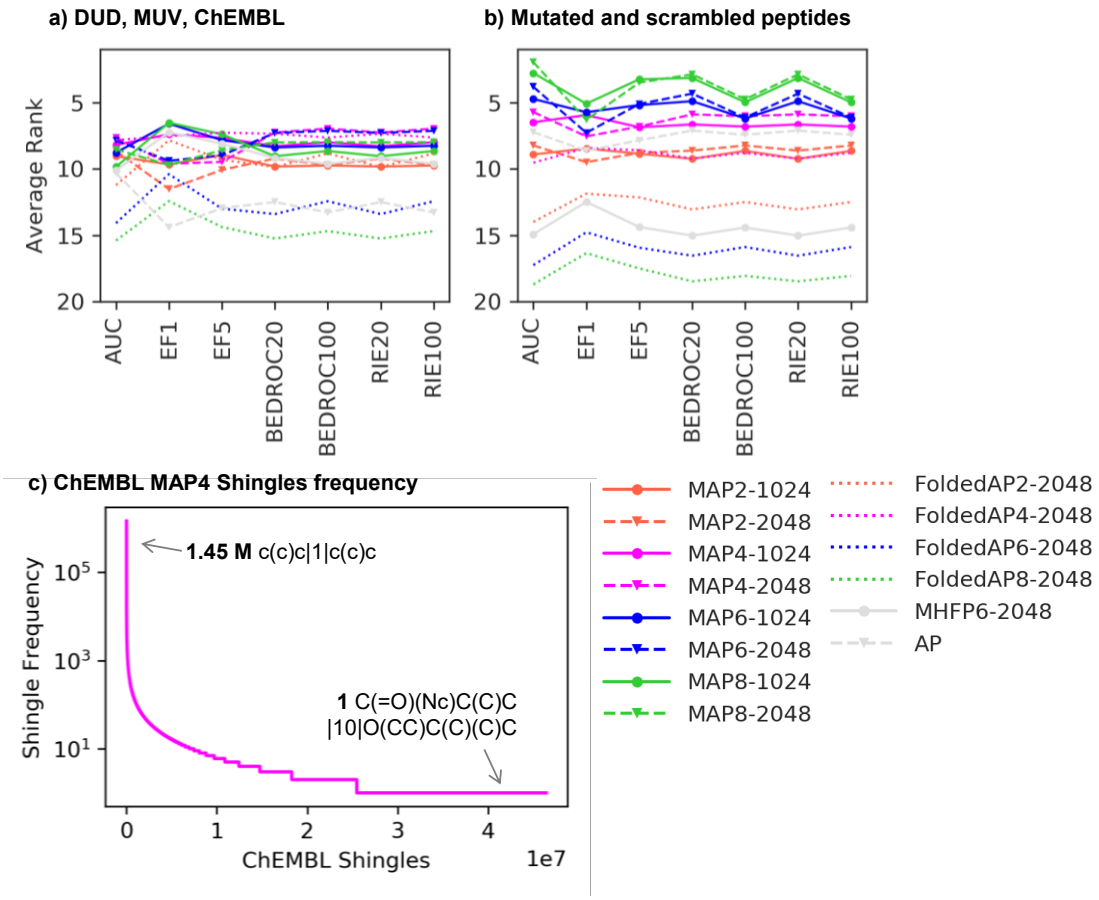


Figure 12. (panels a and b) Average rank of AP2 (orange), AP4 (magenta), AP6 (blue), AP8 (green), in their 1024-dimensions (solid) and 2048-dimensions (dashed) MinHashed implementation (MAPs), and in their 2048-dimensions folded (dotted) implementation (foldedAPs) in the fingerprint benchmark when using only small molecules datasets (17 MUV targets, 21 DUD targets, and 50 ChEMBL targets, a) and only peptide datasets (30 mutated peptide targets and 30 scrambled peptide targets, b). In both panels a and b, MHFP6 (solid) and AP (dashed) are reported in grey. Note that 14 out of 21 fingerprints are shown. (c) ChEMBL MAP4 shingles frequency analysis, examples of shingles with different frequencies are reported.

The above benchmarking study shows that our MinHashed Atom-Pair fingerprint MAP4 performs among the best fingerprints for small molecules and the best fingerprints for peptides, but is the only fingerprint performing best on both benchmarks. We attribute this combined performance to the fact that MAP4 combines circular substructures, which are optimal to describe small molecules, with atom pairs as a method particularly well suited for large molecules. The benchmark among the different MAP versions furthermore shows that the level of detail perceived by the 1024-dimensions MAP4 version is optimal for good performance.

3.3.4 Chemical space maps

To further illustrate the suitability of MAP4 as a molecular fingerprint across various molecule families, we consider different databases covering various molecular size ranges and types (Table 3), and visualize them in form of chemical space tree-maps (TMAPs). [12] These interactive tools can be readily computed exploiting the fact that similarly to MHFP6, MAP4 is a MinHashed fingerprint, for which one can use locality sensitive hashing (LSH) for computing the k-NN tree that is represented in the TMAP even for databases of millions of molecules. The TMAPs discussed below are freely accessible at http://tm.gdb.tools/map4/.

Table 3. Databases illustrated as MAP4 tree-maps.

Database	**Size**[a)]	**HAC**[b)]
ChEMBL[c)]	1,870,343	30.0 ± 17.5
Non-Lipinski ChEMBL	203,850	55.7 ± 38.7
Human Metabolome	114,016	61.7 ± 28.1
SwissProt	9,054	237.4 ± 104.7
DrugBank	229,067	26.2 ± 20.7

a) number of molecules in the database after pre-processing (see methods). b) HAC = heavy atom count given with standard deviation. All non-hydrogen atoms in the molecule. c) The TMAP for ChEMBL is limited to a random 10% subset (187,034 compounds) to reduce server load.

Comparing MHFP6 and MAP4-based TMAPs for the ChEMBL database,[128] its non-Lipinski subset, [14] and DrugBank[127] shows that both fingerprints perform comparably well in organizing these databases. Although one would expect that MAP4 would perform better than MHFP6 in separating molecules by size, this is not the case (Figure 13a/b). The ability of MHFP6 to separate molecules by size reflects the fact that in these databases, large molecules contain either a larger diversity of substructures or simply different substructures compared to small molecules, which results in an implicit size perception in the substructure encoding even if these substructures are small. The ability of both MAP4 and MHFP6 to classify molecules across different size ranges is well illustrated by visualizing phosphorous-containing molecules, which span from inorganic phosphates through cofactors (CoA, NADH) to large

therapeutic oligonucleotides (AGRO100, Figure 13c/d). On the other hand, in TMAPs of the SwissProt dataset MAP4 separates molecules by size much better than MHFP6 (Figure 14a/b). In this case BLAST analogs are also better grouped in the MAP4-based maps than in the MHFP6-based maps, in line with the peptide benchmark study (Figure 14c/d).

MAP4 also performs much better than MHFP6 for mapping the Human Metabolome Database (HMDB). This database contains diverse lipids, phospholipids, carbohydrates, glycosides, amino acid derivatives and more.[17] In this case, MAP4 produces a very well defined TMAP because encoding atom-pairs up to any distance ensures a differentiation between molecules containing different numbers of repetitive substructures such as lipids and glycosides (Figure 15a). By contrast, MHFP6 fails to properly distinguish between related metabolites and the map consists of very large groups of molecules appearing as "grapes" (Figure 15b). Analyzing the occupancy of fingerprint value bins shows that for the three substructure fingerprints, the ten most populated fingerprint value bins contain a large number of molecules, thousands for ECFP4 and MHFP6 and hundreds for TT (Figure 15c). These molecules are lipids and phospholipids, and in the case of ECFP4 and MHFP6, these are the same molecules (Figure 43, Figure 44, Figure 45). By contrast, atom-pair fingerprints contain either a single molecule per bin (MAP4) or at most two or three molecules per bin (AP).

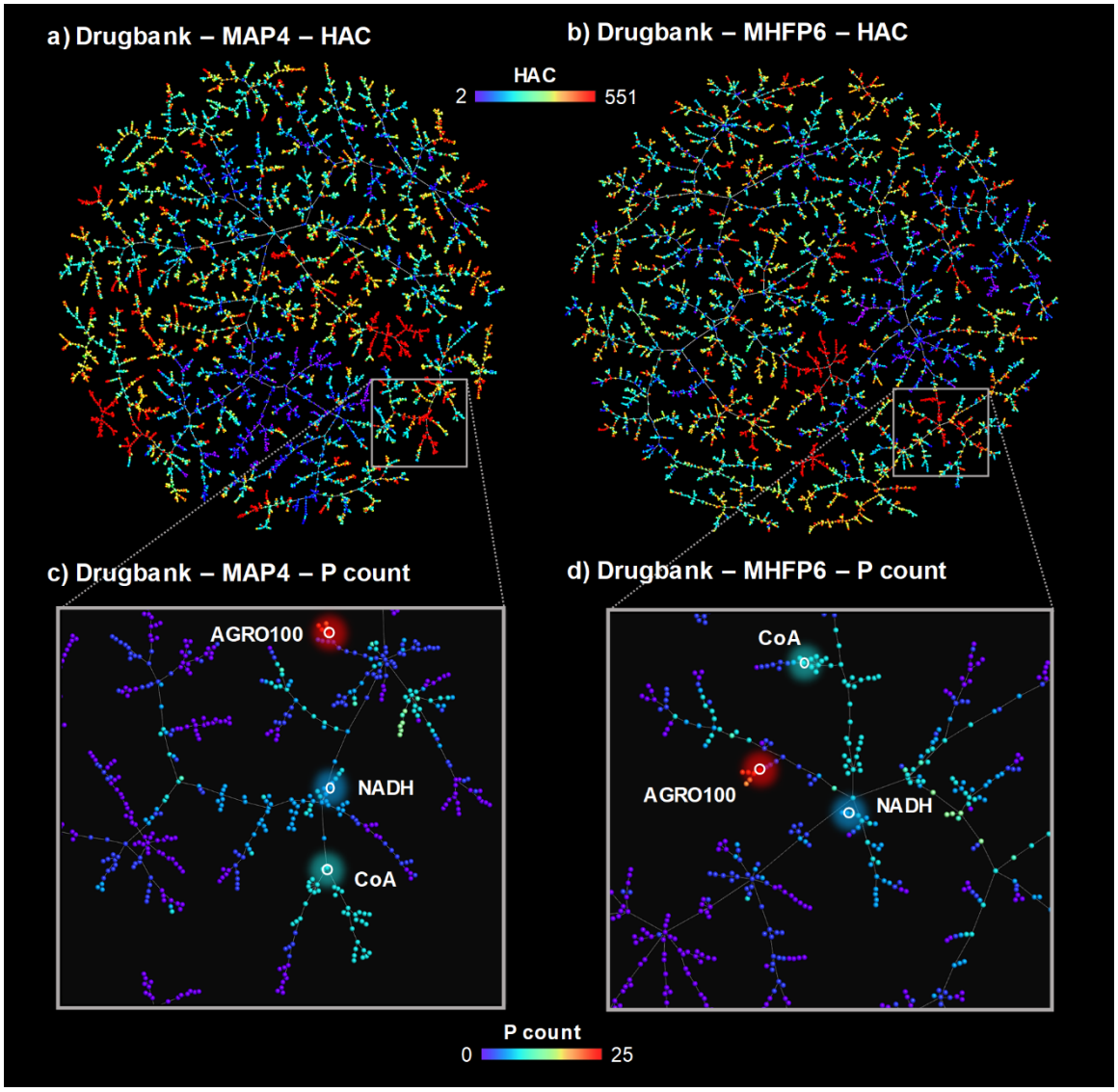


Figure 13. TMAPs of Drugbank using MAP4 and MHFP6. (**a**) MAP4 TMAP color-coded by molecule size (HAC). (**b**) MHFP6 TMAP color-coded by molecule size. (**c**) Close-up view of (**a**) color-coded by the number of phosphorous atoms per molecule (P count). (**d**) Close-up view of (**b**) color-coded by P count. Interactive TMAPs of Drugbank, ChEMBL, and non-Lipinski ChEMBL, color-coded with additional properties, are accessible at http://tm.gdb.tools/map4/.

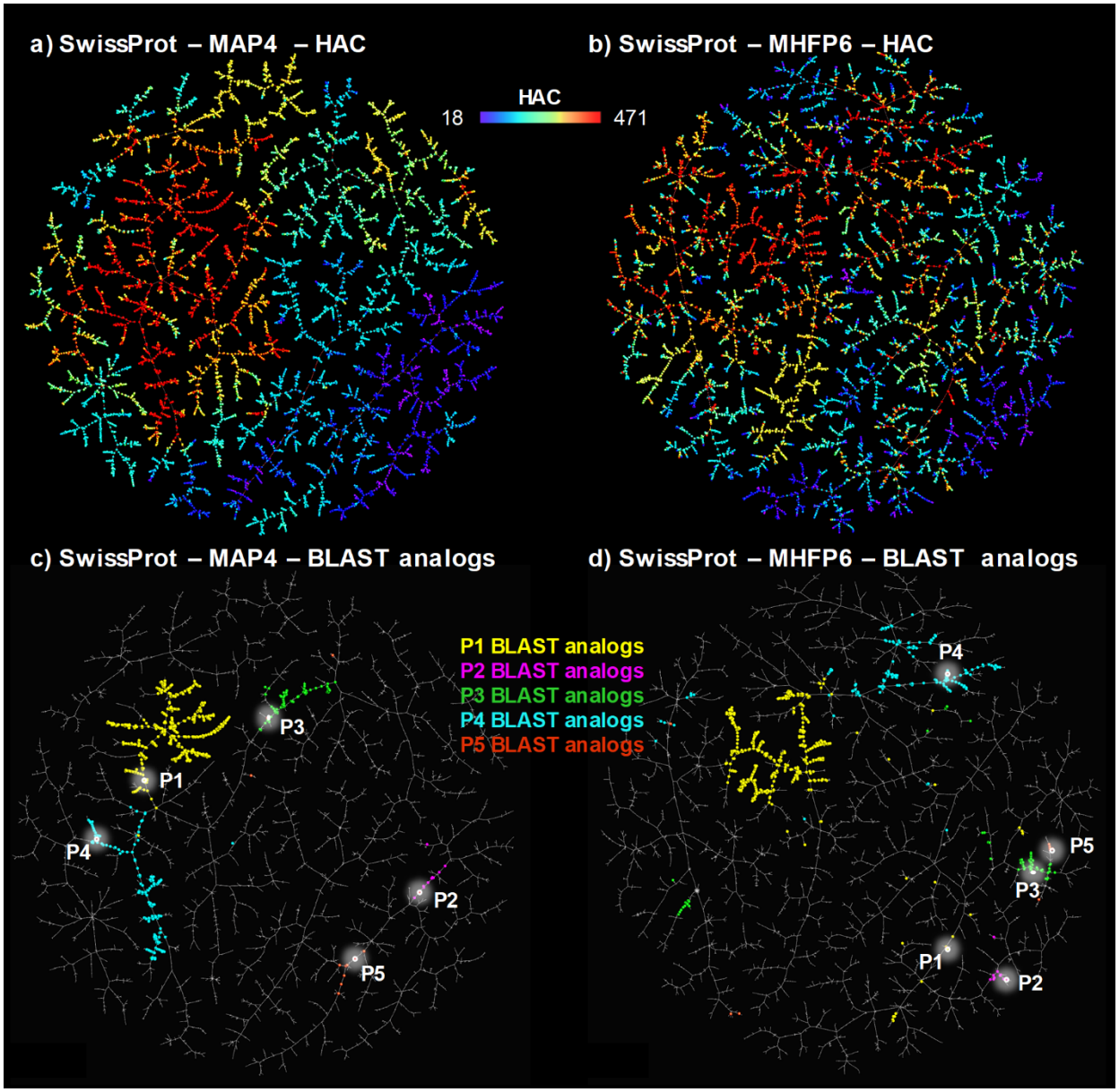


Figure 14. TMAPs of the SwissProt dataset. (**a**) MAP4 TMAP and (**b**) MHFP6 TMAP color-coded by HAC. (**c**) MAP4 TMAP and (**d**) MHFP6 TMAP color-coded by BLAST analogs of MTQRTLRGTNRRRIRVSGFRARMRTASGRQVLRRRRAKGRYRLAVS (P1), MELFAALNLEPIFQLTFVALIMLAGPFVIFLLAFRGGDL (P2), TNRNFLRF (P3), and MRVNITLECTSCKERNYLTNKNKRNNPDRLEKQKYCPRERKVTLHRETK (P4), INLKALAALAKKIL (P5) in the MAP4 (c) and MHFP6 (d) chemical spaces. The interactive maps are accessible at http://tm.gdb.tools/map4/.

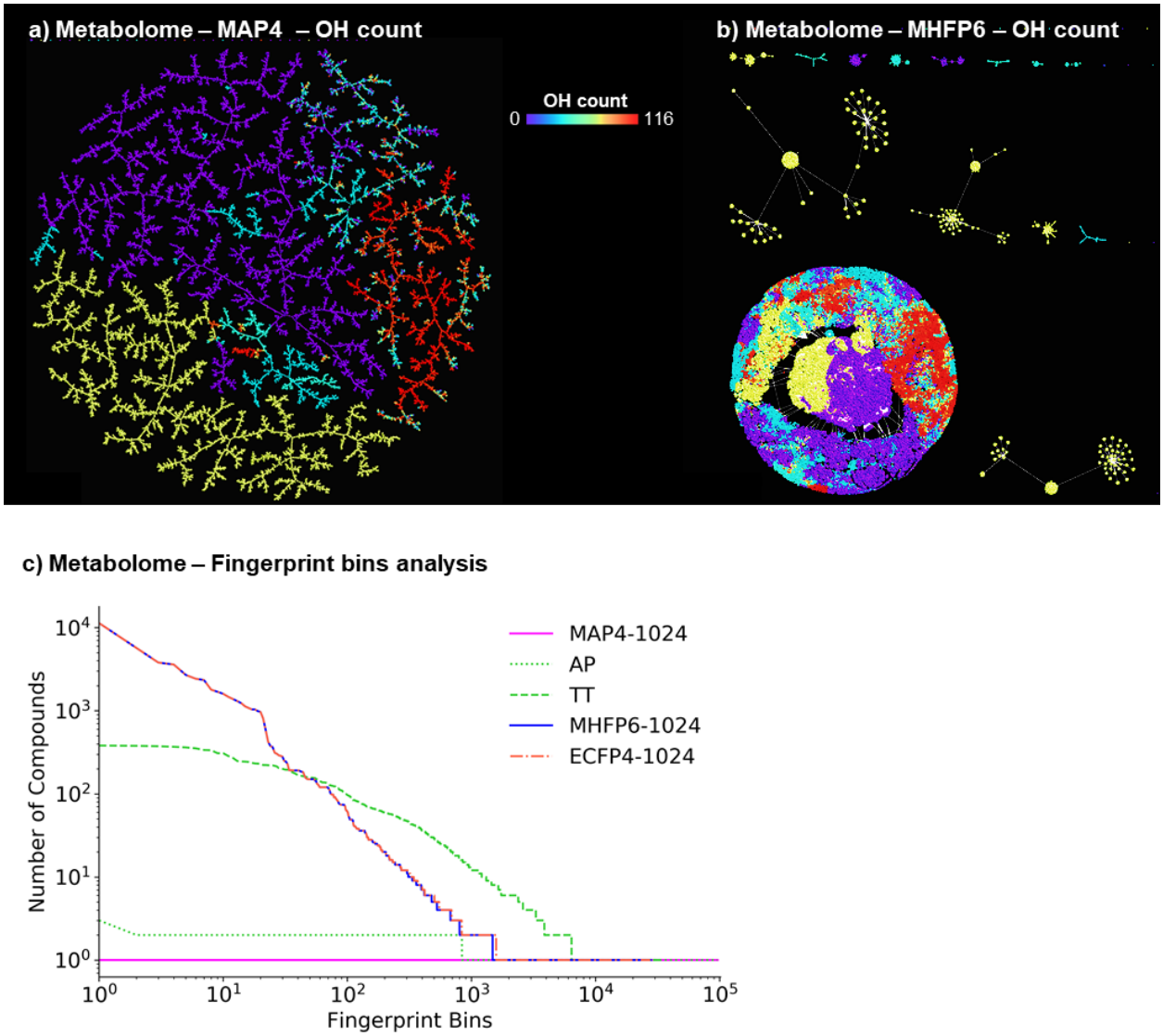


Figure 15. TMAPs of the Human Metabolome Database. (**a**) MAP4 TMAP and (**b**) MHFP6 TMAP color-coded by OH count. The interactive maps with additional properties are accessible at http://tm.gdb.tools/map4/. (**c**) Human Metabolome compounds per fingerprint bins in the MAP4-1024 (magenta line, solid), AP (green line, dashed), TT (green line, dotted), MHFP6-1024 (blue line, solid), and ECFP4-1024 (orange line, dash-dotted) chemical spaces.

3.3.5 Nearest neighbor searches

The difference in the MAP4- and MHFP6-based TMAPs of HMDB reflects the ability of MAP4 to distinguish between closely related metabolites perceived as identical by MHFP6. HMDB contains 96,456 structurally different metabolites not considering stereochemistry. Performing an exhaustive nearest-neighbor (NN) search on these metabolites shows that MAP4

distinguishes all metabolites from one another without exception (Table 4). By contrast MHFP6 finds an indistinguishable NN (JD = 0) in 72.5 % of HMDB molecules. The situation is even slightly worse with ECFP4 (72.9 %) and slightly better with TT (71.1 %). On the other hand, AP sees an indistinguishable NN in only 1,677 molecules (1.7 %) and is therefore almost as good as MAP4.

Table 4. Nearest neighbor analysis of the Human Metabolome database.[a)]

HMBD Subset	All	OH = 0	OH = 1	1 < OH ≤ 4	OH > 4
All	96,456	33,721	10,663	41,493	10,579
JD(MAP4-1024) = 0	0	0	0	0	0
JD(AP) = 0	1,677	13	35	1,611	18
JD(TT) = 0	68,623	27,897	5,782	32,909	2,035
JD(MHFP6-1024) = 0	69,972	28,502	6,215	33,359	1,996
JD(ECFP4-1024) = 0	70,329	28,561	6,243	33,294	2,231

a) Subsets of the Human Metabolome 4.0 Database according to the number of hydroxyl groups per molecule separating lipids (OH = 0, 1) from carbohydrate derivatives (OH > 4). For each subset (column), the number of molecules is indicated in total (All, line 2) and counting those with an indistinguishable nearest neighbor (Jaccard Distance JD = 0) according to the indicated fingerprint (line 3-7). Molecules were considered after removing stereochemical information.

HMDB can be sorted by OH-count, which approximately separates triglycerides and related apolar lipids (OH = 0), diglycerides, alcohols and acids (OH = 1), phospholipids (1 < OH ≤ 4) and carbohydrates (OH > 4). Analyzing the number of indistinguishable NN as a function of OH count shows that AP mostly fails with phospholipid-type molecules (1 < OH ≤ 4), where 96.1 % of the 1,677 AP-indistinguishable NN are found. A remarkable example is provided by the complex phospholipids HMDB0072949 and HMDB0076236, which are distinguished from one another only by MAP4 (Figure 16a). AP also fails to distinguish between 4-phenanthrol (HMDB0059800) and 9-phenanthrol (HMDB0059801), the latter being an inhibitor of the ion channel TRPM4 (Figure 16b).[137] This lack of differentiation by AP is somewhat surprising since all other fingerprints easily distinguish between these two isomers, and reflects the fact that AP is the only fingerprint in the series which does not perceive atom environments but only atomic properties.

MAP4 and AP perceive differences between many closely related metabolites that are indistinguishable for substructure fingerprints. An interesting example among carbohydrates is provided by the branched hexasaccharides HMDB0006605 and HMDB0006614, which only differ from one another by the permutation of the fucoside and 4-sialyl-galactoside at the C(3)-OH and C(4)-OH groups of the central N-acetylglucosamine (Figure 16c). This differentiation is enabled by the encoding of atom-pairs at distances longer than the maximum length spanned by the substructure fingerprints MHFP6 (six bonds), ECFP4 and TT (four bonds).

Encoding atom-pairs at long distances is also what enables atom-pair fingerprints to perform well in the peptide benchmark discussed above where BLAST-analogs must be recovered from scrambled or mutated sequences. This is well illustrated for NN searches in the case of heptapeptides KLLKKLL and KLKKLLL, which are only distinguished from one another by MAP4 and AP (Figure 16d). A similar situation arises when considering oligonucleotides such as the pair ACTG and ATCG which only differ by the permutation of the two central pyrimidine bases (Figure 16e).

Inspecting nearest neighbors of any molecule of interest provides an additional opportunity to explore the content of large databases, often as a means to perform virtual screening to identify analogs. The MinHashed nature of MAP4 enables us to perform extremely rapid approximate nearest neighbor (k-NN) searching using locality sensitive hashing (LSH). We have therefore prepared MAP4 similarity search portals for the ChEMBL, the Human Metabolome, and the SwissProt subset described above, which are freely accessible at http://map-search.gdb.tools/. Note that NN-searches using LSH forests are approximate and not identical with the exact NN-searches using in the benchmarking study, however, it is well-known that the results of approximate k-NN searches based on LSH forests are not significantly different from exact k-NN searches.[138]

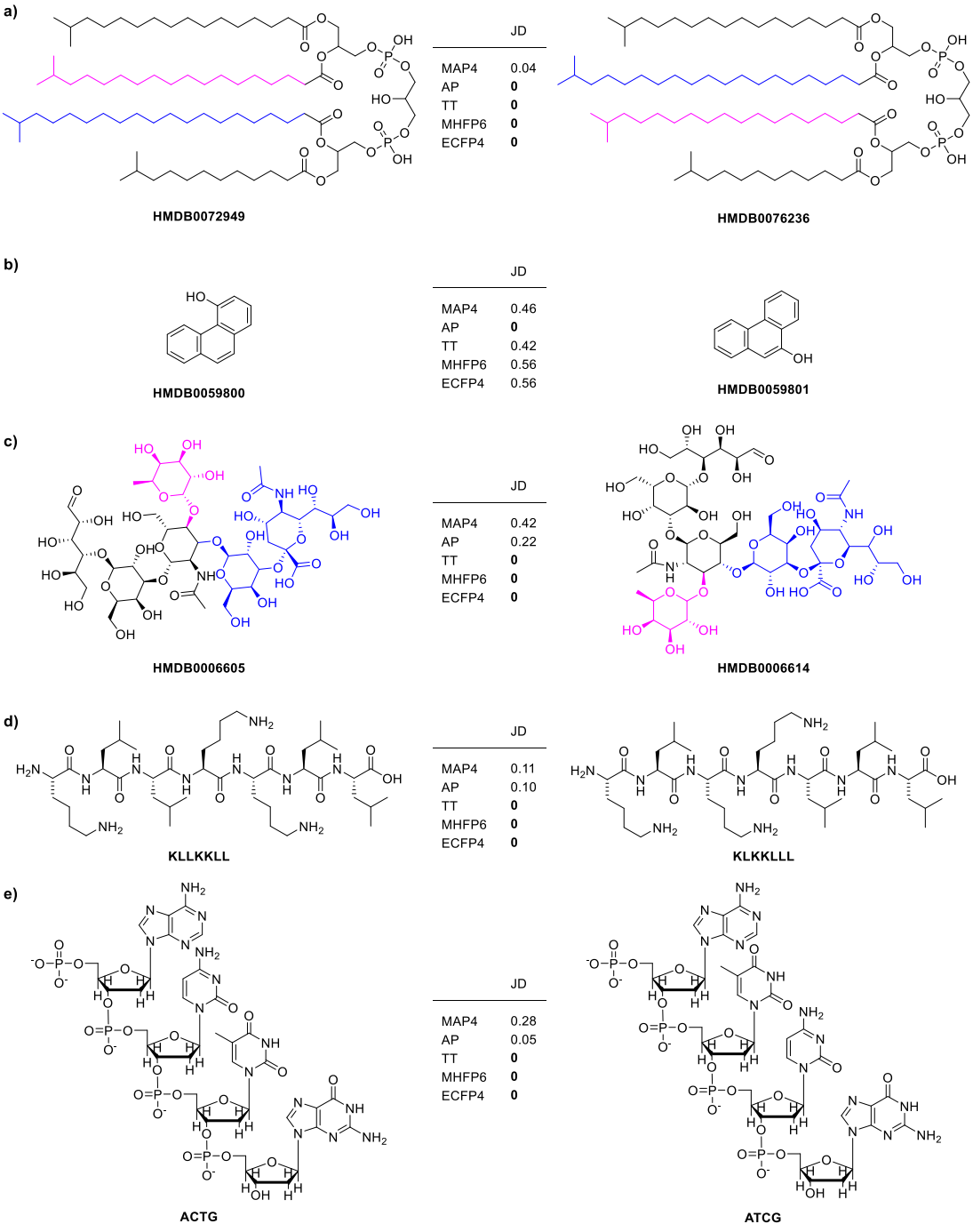


Figure 16. Pairs of molecules better differentiated with MAP4 than with MHFP6, MAP4, TT, AP, and ECFP4 and their JD values. (**a**) Lipids from HMDB, the different position of the lipidic chains is highlighted using blue and magenta. (**b**) Phenanthrol isomers from HMDB. (**c**) Hexasaccharides from HMDB, the α-L-fucosyl and β(3-sialyl)-galactosyl groups exchanged at positions 3 and 4 of the central N-acetylglucosamine are highlighted using blue and magenta (structures as given in HMDB with open-chain form of the first carbohydrate and missing stereochemistry at one center each). (**d**) Scrambled heptapeptides. (**e**) Scrambled tetranucleotides.

3.4 Conclusion

In summary, combining the principles of circular substructures, atom-pairs, and MinHashing produces the MinHashed atom-pair fingerprint MAP4. MAP4 is a new molecular fingerprint performing as good as extended connectivity fingerprints such as ECFP4 and MHFP6 on the Riniker and Landrum small molecule benchmark, and as good as the RDkit AP fingerprint on a new peptide sequence similarity benchmarking set for recovering BLAST analogs among scrambled and mutated peptide sequences, designed to evaluate performance on large molecules. The high performance of MAP4 in the small molecule benchmark is made possible by the substructure encoding which is absent in previous atom-pair fingerprints, while high performance in the peptide benchmark reflects the perception of atom-pairs at unrestricted topological distances which is missing in substructure fingerprints. While the current version of the MAP fingerprint is implemented in Python and therefore it is relatively slow, the performance might increase by rewriting the fingerprint in C or C++.

The MinHashing used for MAP4 allows the construction of k-NN trees and the creation of high-resolution chemical space tree-maps (TMAPs) for databases as diverse as DrugBank, ChEMBL, Swissprot, and the Human Metabolome. The MAP4 based TMAPs are much better defined than those obtained using the substructure MinHashed fingerprint MHFP6, in particular for the case of the Human Metabolome. This is because MAP4 perceives differences among highly similar molecules such as lipids with related fatty acid chains which are not seen by MHFP6. MAP4 also distinguishes between high-similarity pairs of peptides and oligonucleotides perceived as identical by substructure fingerprints such as MHFP6. MAP4 represents a universal fingerprint to search and map the chemical space across molecules of all types and sizes and should be generally useful in the field of cheminformatics.

Chapter Four – Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning

This work is based on the peer-reviewed publication:

Capecchi, A.; Reymond, J.-L. Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning. Biomolecules 2020, 10 (10), 1385. https://doi.org/10.3390/biom10101385.

Abstract

Microbial natural products (NPs) are an important source of drugs. However, their structural diversity remains poorly understood. Here we used our recently reported MinHashed Atom Pair fingerprint with diameter of four bonds (MAP4), a fingerprint suitable for molecules across very different sizes, to analyze the Natural Products Atlas (NPAtlas), a database of 25,523 NPs of bacterial or fungal origin downloaded from https://www.npatlas.org/joomla/. To visualize NPAtlas by MAP4 similarity, we used the dimensionality reduction method tree map (TMAP) (http://tmap.gdb.tools). The resulting interactive map (https://tm.gdb.tools/map4/npatlas_map_tmap/) organizes molecules by physico-chemical

properties and compound families such as peptides and glycosides. Remarkably, the map separates bacterial and fungal NPs from one another, revealing that these two compound families are intrinsically different despite of their related biosynthetic pathways. We used these differences to train a machine learning model capable of distinguishing between NPs of bacterial or fungal origin.

4.1 Introduction

Natural products (NPs) of microbial origin are an important source of drugs. Numerous examples of antibiotic, antifungal, immunosuppressive, anti-inflammatory, and anti-cancer agents on the market originate from fungi or bacteria [139]. A notable effort has been done to explore the known and virtual chemical space of microbial NPs and NPs in general [140–143]. Furthermore, machine learning (ML) has been extensively applied to natural product structures, for example to classify limonoids and protolimonoids [144], to establish the structural class of a natural product with its NMR data [145], to learn estimates of natural product conformational energies [146], to generate derivates of NPs or compounds with natural product characteristics [147–149], to predict Meridian in Chinese traditional medicine [150], and to elucidate the biological effects of natural products [151]. The recently published Natural Products Atlas (NPAtlas) is a collection of 25,523 NPs of fungal and bacterial origin [19]. Among other tools, the NPAtlas website (https://www.npatlas.org/joomla/) provides a global view of the database in a spherical representation. To generate this view, The NPAtlas entries are clustered by Dice similarity [152] using the substructure fingerprint ECFP4 (extended connectivity fingerprint with a diameter of four bonds) [7]. The resulting clusters are grouped in nodes, which are arranged in a spherical plot where the position of each node is determined by molecular formulas. While this representation provides interesting insights on the composition of the NPAtlas, individual compounds cannot be visualized in the global overview but only within clusters. Therefore, comparing compounds across two different clusters is not possible.

A defining feature of NPAtlas is that NPs featured in this database span across a broad range of sizes, with the largest NPs reaching up to almost 3 kDa (Figure 46). We showed recently that the ECFP4 fingerprint, although well suited for small molecule drugs, perform poorly with larger molecules typically found in NP collections such as lipids, oligosaccharides, and peptides [18]. To address this limitation, we recently investigated molecular fingerprints

combining the concept of atom pairs [10], which is well suited to analyze large molecules such as proteins and peptides [6,14,25,98], with extended connectivity substructures and bit compression using MinHash as used in the substructure fingerprint MHFP6 [8], and proposed the MinHashed atom pair fingerprint with a diameter of four bonds (MAP4) as an optimal molecular fingerprint to analyze molecules of very different sizes [18].

Here we asked the question whether analyzing NPAtlas using MAP4 might provide new insights into the composition of this collection. To organize molecules according to their MAP4 similarity, we used TMAP, a recently reported dimensionality reduction method suitable to analyze very large high-dimensional datasets [12]. TMAP performs better for the visualization of large high-dimensional data sets than other dimensionality reduction methods such as t-SNE [153] or UMAP [57]. Furthermore, TMAP is particularly well suited to analyze databases of molecules associated with MinHashed fingerprints.

4.2 Methods

4.2.1 NPAtlas Dataset

The December 2019 version of the NPAtlas was used. This version of the database contains 25,523 entries, 15,759 of fungal origin and 9,764 entries of bacterial origin, with no entry sharing bacterial and fungal origin. For each compound, simplified molecular-input line-entry system (SMILES), molecular weight (MW), origin (fungal or bacterial), and the DOI of the associated publication were downloaded. For the MAP4 fingerprint calculation the SMILES were canonicalized [131] and the stereochemistry was removed using the RDKit toolkit [123]. After removing stereochemistry, the NPAtlas counts 23,928 unique SMILES and 76 entries common among both origins.

4.2.2 MAP4 Fingerprint

The MAP4 fingerprint combines the circular substructure and atom pair fingerprints concepts. MAP4 encodes each atom pair in a molecule as the SMILES of the circular substructure of radii 1 and 2 around both atoms and the distance in bonds that separates them. The resulting set of strings is hashed to integers using the SHA-1 algorithm [40] and MinHash scheme [154]. The obtained MAP4 fingerprint is an array of unsorted numbers, where each feature is characterized by its value and its position in the array (index). MAP4 perceives substructure details while maintaining a global overview; therefore, it is suitable to describe molecular structures across different sizes. The similarity between two MAP4 fingerprints *a* and *b* is calculated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the number of elements of fingerprint *a*. The similarity between two MinHashed MAP4 fingerprints calculated as described above is an estimation of the Jaccard Similarity between the two non-MinHashed objects [154]. For a detailed explanation if the MAP4 implementation and benchmark please refer to our recent publication [18]. The 1024-dimensions MAP4 fingerprint of all NPAtlas entries was calculated using canonical SMILES without stereochemistry information.

4.2.3 TMAP Layout

The TMAP layout was calculated from the MAP4 fingerprint dataset using the open source implementation of TMAP[12]. In short, the indices generated by the MinHash procedure of the MAP4 calculation are used to create a locality-sensitive hashing (LSH) forest [42] of *n* trees. For each NPAtlas entry, the *k* approximate nearest neighbors (NNs) in the MAP4 feature space are then extracted from the LSH forest to form a graph in which nodes are the structures and edges are the NN relationships weighted by the fingerprint distance. The Kruskal's algorithm is then applied to remove cycles and to find the path with the lowest total distance between all

molecules in the graph [60]. Finally, Fearun [24] is used to interactively display the obtained minimum spanning tree. In this this study we set n = 32 and k = 20.

4.2.4 Properties Calculation

For all NPAtlas entries, the number of hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD), logP following Crippens approach (AlogP) [155], topological polar surface area (TPSA), and fraction of sp3 carbon (fsp3C) were calculated with RDKit. The boiling point was calculated using the open source code of the JRgui [156] as the Joback boiling temperature (T_{Job}, Equation 5) [157],

$$T_{Job} = 198.2 + \sum_i N_i t_{bi} \qquad \text{Equation 5}$$

where and N_i is the occurrence of a functional group in the molecule and t_{bi} is its empirically obtained contribution value. Molecules that violated more than one Lipinski rules [13] were labelled as non-Lipinski. To identify glycosylated and/or peptidic structures, Daylight [158] SMARTS language was used. SMILES arbitrary target specification (SMARTS) were used with RDKit to identify NPAtlas entries containing a dipeptide substructure, defined as "[NX3,NX4+][CH1,CH2][CX3](=[OX1])[NX3,NX4+][CH1,CH2][CX3](=[OX1])[O,N]", or a glycoside substructure, defined as "[CR][OR][CHR]([OR0,NR0])[CR]".

4.2.5 TMAP color gradients

The calculated properties were used to color the generated TMAP. For a clearer color gradient, some of the highest and lowest displayed values of the non-ranked properties have been adjusted. All MW values ≥ 1,000 Da are displayed as 1,000 Da, all boiling point values ≥ 2,000 K are displayed as 2,000 K, all HBD count values ≥ 10 are displayed as 10, all AlogP values ≥ 8 are displayed as 8, all AlogP values ≤ -2 are displayed as -2, and all TPSA values ≥ 500 are displayed as 500. The color-codes of the ranked property values were obtained by average

ranking them using SciPy [133]. In average ranking if two or more values have the same rank the average rank of the tied values is assigned to each of them. For details on TMAP please refer to the related publication [12].

4.2.6 Support vector machine (SVM) and *k*-nearest neighbor (*k*-NN) classifiers

The *k*-nearest neighbor (*k*-NN) algorithm is a simple ML method that predicts the query to belong to the class most found amongst its *k* nearest neighbors. A support vector machine (SVM) represents a more complex ML approach; an SVM maps its input into a high-dimensional feature space and tries to find the best separation between two classes, such as they entirely lay on the opposite side a hyperplane. To do so, the SVM maximizes the margin between the closest points, known as support vectors, and the hyperplane. Mapping features explicitly into a higher dimensional space is computationally expensive and not feasible even for small datasets. To avoid it the SVM uses the so-called "kernel trick", which essentially uses a similarity matrix of the input data instead of the input itself; this allows the SVM to define the hyperplane and the support vectors in a less expensive manner [159]. In cheminformatics, both *k*-NN and SVM inputs can range from SMILES to various molecular descriptors. For this work three classifiers were implemented: a MAP4 based *k*-NN (MAP4 k-NN), a MAP4 based SVM (MAP4 SVM), and an SVM based on physico-chemical properties (physchem SVM).

The MAP4 SVM and MAP4 k-NN classifiers were implemented as follows. The canonicalized SMILES without stereochemistry information used to generate the TMAP were made unique, and they were assigned to training or test set with a 50% random split. The 35 unique SMILES of the 76 entries common between both origins were randomly assigned to one origin. Both classifiers were trained using MAP4 fingerprints. In both cases the class weights were inversely proportional to the class frequency, and their hyperparameter were optimized using a 5-fold cross validation. During the 5-fold cross validation, 20% of the

training set was left out as validation set, and the final set of parameters maximized the ROC AUC on the validation set. For the SVM classifier the hyperparameter C was optimized among the values 0.1,1, 10, 100, and 1000, resulting in C = 10. The SVM utilized a custom kernel that calculates the similarity matrix between two MAP4 fingerprints. Platt scaling [160] was used to obtain probabilistic prediction values. For the *k*-NN model the number of nearest neighbors *k* was optimized among the values 5, 7, 9, and 11, resulting in $k = 7$. As a distance metric between two MAP4 fingerprints the *k*-NN classifier uses one minus the similarity between the two fingerprints.

The physchem SVM model was trained with the same training/test split, but using the MW, fsp3C, HBA, HBD, AlogP, TPSA, and calculated boiling point as input. The properties were scaled to zero mean and unit variance. A radial basis function (RBF) kernel [161] was used, and the hyperparameters C and γ were optimized with a grid search among. C was optimized considering the values 0.1,1, 10, 100, and 1000, resulting in C= 10, and γ was optimized considering the values 0.01, 0.1, 1, 10, and 100, resulting in $\gamma = 1$.

For the evaluation of the classifiers we considered the class "bacterium" to be the positive class and the class "fungus" to be the negative one. All SVM and the *k*-NN classifiers were implemented using scikit-learn [162], and all not mentioned hyperparameters were used in their default values. The source code for all classifiers can be found at https://github.com/reymond-group/MAP4-Chemical-Space-of-NPAtlas.

4.2.7 Classifiers evaluation metrics

ROC AUC is the area under the ROC curve, and the ROC curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR):

$$TPR = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{TP + FP}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives predicted by the classifier.

The F1 score is defined as the harmonic mean of precision and recall:

$$Precision = TPR$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

The balanced accuracy is defined as:

$$Balanced\ accuracy = \frac{TPR + \frac{TN}{TN + FN}}{2}$$

The Matthews correlation coefficient (MCC) is a correlation between the observed and the predicted class and it is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In all metrics, the probabilistic prediction values were converted into binary classification values using a threshold of 0.5.

4.3 Results and discussion

4.3.1 The TMAP of NPAtlas

The 25,523 structures in NPAtlas were downloaded and encoded using the MAP4 fingerprint, which is well suited to analyze molecules across different sizes such as those in NPAtlas ranging between 70 and 2,900 Da in MW (Table 5, Figure 46, method sections 4.2.1 and 4.2.2).

The generated dataset was then visualized using TMAP which represents the minimum spanning tree connecting nearest neighbors, here according to the MAP4 similarity measured as Jaccard distance (Figure 47, see method section 4.2.3 for details). To understand how the NPs in NPAtlas are organized on the MAP4 TMAP, we generated color codes based on various physico-chemical descriptors, as well as on categorical classification by compound type and observed or predicted origin (Table 5, method section 4.2.4 and 4.2.5, Figure 48, Figure 49, Figure 50, https://tm.gdb.tools/map4/).

Inspecting the colored TMAPs reveals that molecules are organized by structural features. For example, inspecting the TMAP colored by MW shows that most of the high MW compounds (MW ≥ 1,000 Da, 6.8% of NPAtlas) belong to three structural families, namely peptides type compounds (minimal substructure: dipeptide), glycosides (minimal substructure: cyclic N- or O-acetal) and glycopeptides (both substructures present) (Table 6, Figure 17a, Figure 46). Typical examples of such large NPs are shown in Figure 18, featuring the cyclic peptides jizanpeptin A (NPA022688, bacterial) [163] and arbumelin (NPA020152, fungal) [164], the glycosides butirosin A (NPA009292, bacterial) [165] and quinofuracin A (NPA005440, fungal) [166], and the glycopeptides cycloaspeptide F (NPA000712, the only fungal glycopeptide in NPAtlas) [167] and orienticin D (NPA021348, bacterial) [168].

Another striking insight is provided by inspecting the TMAP colored by the fraction of sp3 carbons (fsp3C, Figure 17b), which allows the identification of areas rich in aromatic polyphenols with very low fsp3C such as nocatrione A (NPA014210, bacterial)[169] and sydowiol E (NPA001030, fungal) [170], as well as areas populated by terpenoids with very high fsp3C such as neoverrucosane diterpenoids (e.g. neoverrucosan-5β,9β,18β-triol, NPA001820, bacterial) [171] and the anti-influenza virus diterpene wickerol B (NPA008911, fungal) [172]. The structures of these compounds are shown in Figure 18.

The TMAP not only organizes molecules by structural features, but also separates molecules according to their origin, with fungal and bacterial NPs forming well-defined groups across the TMAP (Figure 17c). This separation is striking because biosynthetic pathways in bacteria and fungi are generally similar, and because the different compound families contain NPs of both bacterial and fungal origin (Table 6).

Table 5. Calculated properties of NPAtlas molecules available as TMAP color-codes.

Property	Min. value	Max. value	25% quantile	50% quantile	75% quantile
Molecular weight [a)]	70.1	2,901.3 (1,000 [f)])	292	408.9	562.6
Sp3 C fraction [a)]	0.0	1.0	0.4	0.6	0.7
HBA count [a,b)]	0	68 (20 [f)])	4	6	9
HBD count [a,c)]	0	47 (10 [f)])	3	2	4
AlogP [a,d)]	-28.9 (-2 [g)])	33.8 (8 [f)])	1.2	2.5	4.1
TPSA [a,e)]	0.0	1,135.81 (500 [f)])	69.64	99.66	152.8
Boiling point [a, h)]	311.5	7,806.5 (2,000 [f)])	890.8	1,141.6	1,518.5
Is Lipinski	Categorical: yes/no				
Substructures [i)]	Categorical: contains dipeptide moiety/contains glycoside moiety/contains dipeptide and glycoside moieties				
Origin	Categorical: Bacterial/Fungal				
MAP4 SVM[j)] prediction	Categorical: Bacterial/Fungal				
MAP4 SVM[j)] performances	Categorical: correct/wrong				

a) Continuous properties; shown also as rank in the map. b) Hydrogen bond acceptors (HBA). c) Hydrogen bond donors (HBD). d) LogP Calculated following Crippens approach (AlogP). e) topological polar surface area (TPSA). f) Maximum value shown in the map, all values above are represented to with the same color code. g) Minimum value shown in the map, all values below are represented to with the same color code. h) Joback calculated boiling point. i) SMARTS matched substructures. j) Support vector machine (SVM).

Table 6. NPAtlas entries and unique publications number according to origin and molecular weight.

	Fungal[a)]	Bacterial[a)]
NPAtlas entries (≥ 1,000 Da)	15,759 (347)	9,764 (1,392)
Unique publications[b)]	6,110 (145)	4,653 (711)
Peptides (≥ 1,000 Da)[c)]	722 (311)	2,144 (901)
Glycosides (≥ 1,000 Da)[d)]	814 (12)	1,616 (421)
Glycopeptides (≥ 1,000 Da)[e)]	1 (0)	112 (89)
Aromatic NPs (≥ 1,000 Da)[f)]	1,322 (0)	800 (31)
Aliphatic NPs (≥ 1,000 Da)[g)]	2,184 (59)	1,366 (220)

a) Natural product origin. b) Number of unique publications used for the extraction of all NPAtlas entries c) Containing a dipeptide moiety. d) Containing a glycoside moiety. e) both glycoside and dipeptide moiety. f) fsp3C < 0.2. g) fsp3C > 0.8.

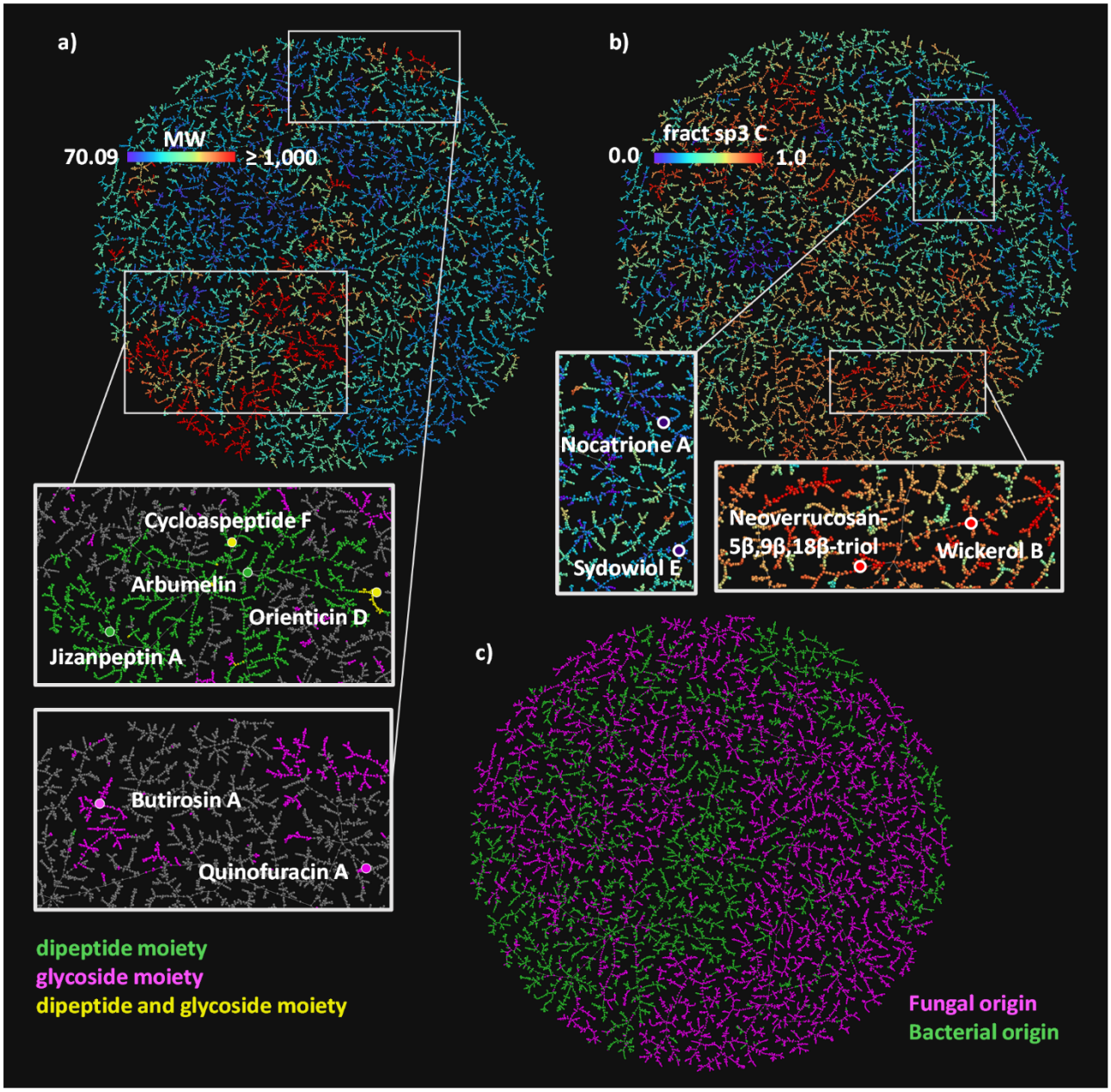


Figure 17. (a) NPAtlas MAP4 TMAP colored by MW, with a rainbow scale where the lowest values are purple, and the highest values are red. Two areas of the map are zoomed and colored by SMARTS substructure match: compounds containing a dipeptide moiety are highlighted in green, compounds containing a glycoside moiety are highlighted in magenta, compounds containing both moieties are highlighted in yellow; six examples of NPAtlas entries are reported with the same color code. (b) The NPAtlas MAP4 TMAP colored by fsp3C with a rainbow scale where the lowest values are purple, and the highest values are red. A low and a high fsp3C area of the map are zoomed, and two examples of polyphenols and of terpenoids are reported. (c) The NPAtlas MAP4 TMAP colored by a microbial origin classification, the compounds originated from fungi are colored in magenta, the compounds produced by bacteria are colored in green.

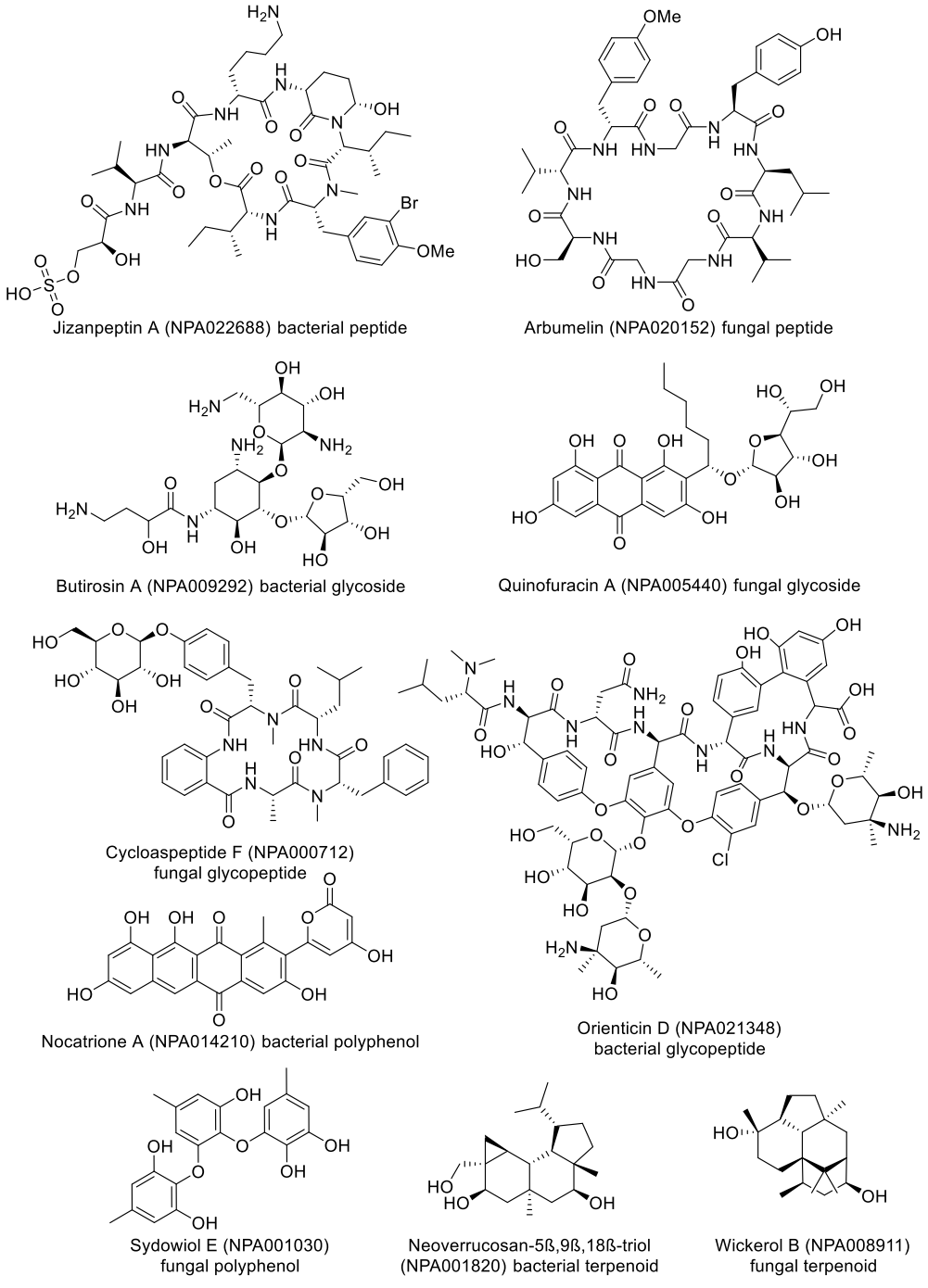


Figure 18. Structural formula of natural products examples selected from the TMAPs in **Figure 17**.

4.3.2 Distinguishing between bacterial and fungal NPs

The separation between bacterial and fungal NPs on the MAP4 TMAP and the fact that the map also separates NPs by physico-chemical descriptor values suggested to us that ML models trained either with the MAP4 fingerprint or simply with physico-chemical descriptors might be able to distinguish between NPs of bacterial or fungal origin. We investigated SVM and *k*-NN models since this type of ML models are generally well suited for classifying bioactive molecules [173]. We considered both an SVM and a *k*-NN model with MAP4, and only an SVM model with physico-chemical descriptors and we evaluated their performance on the test set (see section 4.2.6).

The MAP4 SVM was the best performing model with an area under the receiver operating characteristic curve (ROC AUC) of 0.97, an F1 score of 0.91, a balanced accuracy of 0.93, and a Matthews correlation coefficient (MCC) of 0.86 (Table 7). The MAP4 *k*-NN classifier also had excellent evaluation metrics with an accuracy of 0.90 and an MCC of 0.8, suggesting the high performance of the MAP4 SVM classifier might depend on a nearest neighbor effect. On the other hand, the physchem SVM performed significantly worse than the MAP4 based classifiers and was only partially capable of distinguishing between bacterial and fungal NPs (F1 score and a balanced accuracy above 0.7). This suggests that successful classification requires a model distinguishing between specific substructures and not only overall molecular properties. For closer inspection, the prediction (fungal or bacterial origin) and the performance (correct or wrong) of the best performing classifier (MAP4 SVM) are color-coded on the MAP4 TMAP of NPAtlas (Figure 50).

Table 7. SVM and *k*-NN classifiers performance on the test set.

Classifier	ROC AUC[a)]	F1 score[a)]	Balanced accuracy[a)]	MCC[a)]
MAP4 SVM[b)]	0.97	0.91	0.93	0.86
MAP4 *k*-NN[c)]	0.96	0.88	0.90	0.81
Physchem SVM[d)]	0.86	0.73	0.78	0.56

a) Area under the receiver operating characteristic curve (ROC AUC), F1 score, balanced accuracy, and MCC are metrices used to evaluate a machine learning model. MCC can assume values from -1 to 1, all other parameters can assume values from 0 to 1, and in all cases 1 is a perfect classification. b) SVM classifier trained with the MAP4 fingerprint. c) *k*-NN classifier trained with the MAP4 fingerprint. d) SVM trained with physiochemical properties.

4.3.3 Predicting the origin of newly discovered NPs

Discussion with natural product chemists informed us that assigning NPs to their origin only from its chemical structure is not trivial, and can be problematic when isolating a new NP due to the occurrence of endosymbiosis, i.e. the fact that bacteria often live as symbionts within larger organisms [170,174]. We therefore asked the question whether our MAP4 SVM classifier would correctly predict the origin of NPs newly reported in 2020 and which are not part of NPAtlas (Table 8). To our delight, the classifier correctly predicted the fungal origin for the newly reported epicospirocins 1 [175], penicimeroterpenoid A [176], and rhizolutin [177], as well as the bacterial origin of the recently reported bosamycin A [178]. The correct origin assignment is probably related to the presence of structurally similar NPs within the NPAtlas training set, illustrated here by the MAP4 nearest-neighbor NPs aspermicrone A [179], isocitreohybridone H [180], Monacolin K [181], and AIP I [182] (Figure 19).

When challenged with the recently reported NP phakefustatin A isolated from the marine sponge *Phakellia fusca* [183] (Figure 19), the MAP4 SVM classifier predicted a bacterial origin (Table 8). Indeed, the NPAtlas training set contained closely related NPs of bacterial origin such as the MAP4 NN Samoamide A [184] (Figure 19). Although phakefustatin A was isolated from a marine sponge, our prediction is probably correct because many marine sponges contain endosymbiotic bacteria, which can make up to 60% of the sponge biomass and are

often responsible for the production of metabolites [185]. More specifically, it is known that *Phakellia fusca* coexists with diverse actinobacteria which have been held responsible for the production of many bioactive NPs found in the sponge [186].

Table 8. MAP4 SVM classification of new microbial natural products and of Phakefustatin A.

Natural Product	MAP4 SVM[a)] fungal, bacterial	Training set nearest neighbor (NN)	JD from NN[b)]
Epicospirocin 1	**0.99**, 0.01	Aspermicrone A (NPA024935)	0.66
Penicimeroterpenoid A	**1.0**, 0.0	Isocitreohybridone H (NPA016454)	0.63
Rhizolutin	**0.83**, 0.17	Monacolin K (NPA009354)	0.80
Bosamycin A	0.04, **0.96**	AIP I (NPA010987)	0.77
Phakefustatin A	0.12, **0.88**	Samoamide A (NPA022212)	0.68

a) Predicted origin: fungal or bacterial. b) Approximated Jaccard Distance (JD, see methods for details) from the training set NN.

While the example above might be a case of endosymbiosis and potential origin misclassification, it must be noted that our MAP4 SVM classifier can only label NPs as of bacterial or fungal origin. In fact, our classifier mistakenly assigns such classification to well-known non-microbial NPs (Table 23, Figure 51). An extension of our analysis to non-microbial natural products could be of interest, however the task cannot be completed due to a lack of annotated public datasets for NPs of diverse origins [187,188].

4.4 Conclusion

In summary, we showed that mapping the 25,523 NPs reported in NPAtlas as a MAP4 TMAP organizes molecules by physico-chemical properties and by substructures and thereby provides an unprecedented insight into the composition of this collection. Most strikingly, the map separates the different NPs according to their bacterial or fungal origin. Furthermore, a SVM model trained with the MAP4 fingerprint dataset performs remarkably well in distinguishing between fungal and bacterial NPs. The classifier can be of aid where the origin of a natural product is unknown, especially when the molecule is isolated from a symbiotic complex. The

MAP4 TMAP of NPAtlas is accessible at https://tm.gdb.tools/map4/ and the source code is available at https://github.com/reymond-group/MAP4-Chemical-Space-of-NPAtlas.

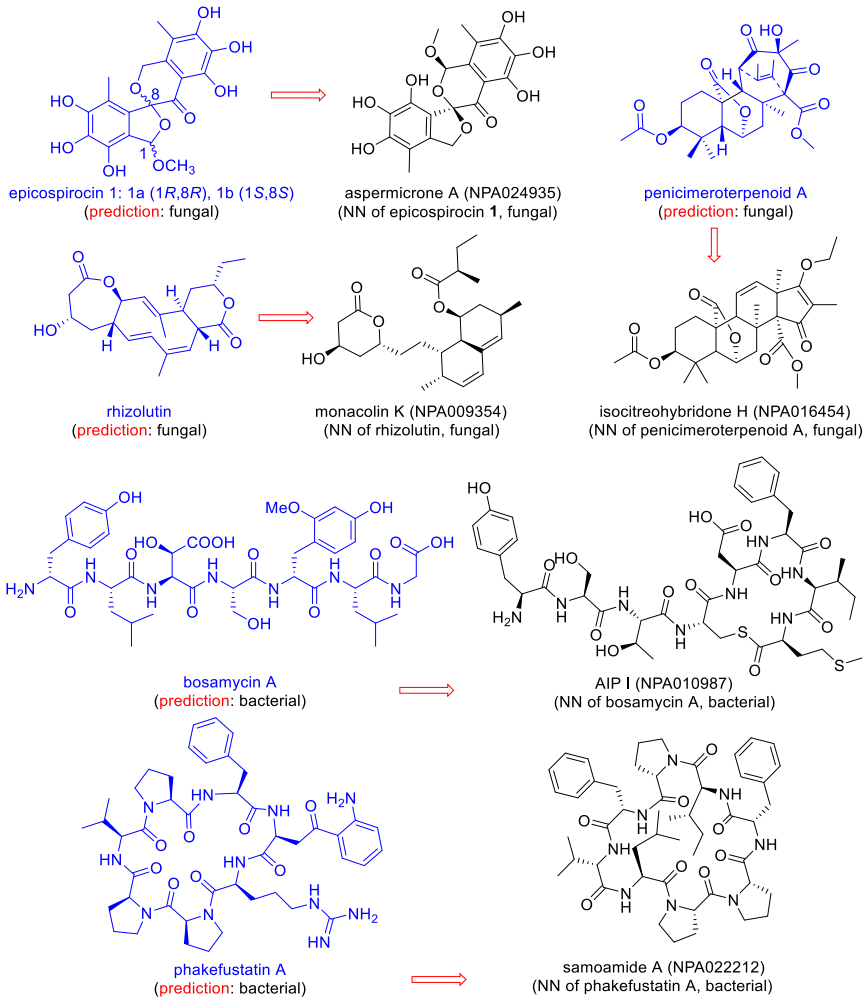


Figure 19. Examples of natural products reported in 2020, absent from NPAtlas, annotated with their predicted origin, and connected to its MAP4 NN in the training set.

Chapter Five – Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database

This work is based on the ChemRxiv preprint:

Capecchi, A.; Reymond, J.-L. Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database. 2021. https://doi.org/10.33774/chemrxiv-2021-gxjgc

Abstract

Natural products (NPs) represent one of the most important resources for discovering new drugs. Here we asked whether NP origin can be assigned from their molecular structure in a subset of 60,171 NPs in the recently reported Collection of Open Natural Products (COCONUT) database assigned to plants, fungi, or bacteria. Visualizing this subset in an interactive tree-map (TMAP) calculated using MAP4 (MinHashed atom pair fingerprint) clustered NPs according to their assigned origin (https://tm.gdb.tools/map4/coconut_tmap/), and a support vector machine (SVM) trained with MAP4 correctly assigned the origin for 94% of plant, 89% of fungal, and 89% of bacterial NPs in this subset. An online tool based on an SVM trained with the entire subset correctly assigned the origin of further NPs with similar

performance (https://np-svm-map4.gdb.tools/). Origin information might be useful when searching for biosynthetic genes of NPs isolated from plants but produced by endophytic microorganisms.

5.1 Introduction

Due to the importance of natural products (NPs) in drug discovery,[189,190] there is a considerable interest in describing and understanding their structural diversity, particularly by exploiting NP databases[140] using *in silico* methods such as machine learning (ML).[188] Computational approaches have been reported to distinguish between NPs and non-Nps,[191–194,147] between terrestrial and marine NPs,[195] and to classify NP structural types[196,197] and visualize their chemical space.[67]

In our own approach to this problem,[21] we recently analyzed NPAtlas, an open-access database listing 25,523 NPs from bacterial or fungal origin,[19] by computing the MAP4 fingerprint (MinHashed Atom-Pair fingerprint up to four bonds)[18] of each NP and creating a TMAP (tree-map) [12] of the resulting high-dimensional dataset. In this analysis, NPs from bacterial or fungal origin formed separated clusters. This separation effect was confirmed by showing that a support vector machine (SVM) trained with the MAP4 of NPAtlas was able to distinguish bacterial or fungal origin, including a recently reported NP isolated from the marine sponge *Phakellia fusca* assigned by our classifier to be of bacterial origin, in line with the fact that many NPs from this sponge originate from endosymbiotic actinobacteria.[183,186]

The possibility to assign the origin of NPs from their structure was intriguing because most NPs are secondary metabolites produced by biosynthetic gene clusters[198] which are sometimes transferred between different organisms.[199] Such horizontal gene transfer may reflect adaptative relationships between host organisms such as plants and sponges and endosymbiotic bacteria or fungi.[200] Among the many endophytic NPs,[201,202] striking examples include the cancer drug paclitaxel, a plant NP also produced by endophytic fungi of the yew tree,[203,204] and maytansine, used in antibody-drug conjugates for cancer therapy and produced by endophytic bacteria in plants.[205] Due to the very widespread occurrence of endophytic

bacteria and fungi in plants, we asked whether our MAP4 analysis might be able to distinguish plant NPs from bacterial and fungal NPs. To test this hypothesis, we considered the recently reported COCONUT database, a recently reported open-access database currently offering the most extensive coverage and including plant NPs.[20]

5.2 Results and discussion

5.2.1 Chemical space analysis of plant and microbial NPs from the COCONUT database

COCONUT collects over 400 thousand NPs from 52 different databases, 135 thousand of which are annotated with a taxonomical origin. For our analysis, we considered the 68 thousand entries annotated with a source organism that were also associated with a publication. We focused on those annotated as originating from plants (50 %), fungi (23 %), or bacteria (16 %), leaving out a smaller subset of NPs originating from animals (2 %), homo sapiens (2.5 %), of marine origin (1.5 %), or lacking a superclass annotation (5 %). The selected subset of 60,171 NPs contained 33,772 plant NPs, 15,648 fungal NPs and 10,751 bacterial NPs.

The subset spanned from molecular weight MW = 81 Da for 1,2-dihydropyridine, a plant NP,[206] to MW = 2,901 Da for lacticin 481, a bacterial peptide.[207] Plant NPs dominated the intermediate molecular weight range ($200 < MW < 800$), while fungal NPs were most abundant in the low molecular weight range ($MW \leq 200$) and bacterial NPs in the high MW range ($MW \geq 800$). The three series had rather similar distributions of the fraction of sp^3 carbon atoms (Fsp3), which measures the degree of saturation. However, the estimated octanol:water partition coefficient AlogP indicated that highly polar NPs were almost absent from fungal NPs. Furthermore, plant NPs had overall higher percentages of glycosides, while peptides were almost absent from plant NPs and most abundant in bacterial NPs (Table 9).

Table 9. Property distribution and origin of the 60,171 COCONUT entries with a DOI and annotated as plants, fungal, or bacterial.

	Plants NPs[a)]	Fungal NPs[a)]	Bacterial NPs[a)]
MW ≤ 200[b)]	7,072 (21%)	4,919 (31%)	2,237 (21%)
200 ≤ MW < 80 [b)]	24,078 (71%)	10,111 (65%)	6,066 (56%)
MW ≥ 800[b)]	2,622 (8%)	618 (4%)	2,448 (23%)
Fsp3 ≤ 0.2[c)]	4,213 (13%)	1,580 (10%)	1,073 (10%)
0.2 ≤ Fsp3 < 0.8[c)]	22,032 (65%)	11,334 (72%)	7,986 (74%)
Fsp3 ≥ 0.8[c)]	7,527 (22%)	2,734 (18%)	1,692 (16%)
AlogP ≤ -2 [d)]	4,855 (14%)	373 (2%)	1,446 (13%)
-2 ≤ AlogP < 8 [d)]	28,315 (84%)	15,000 (96%)	8,906 (83%)
AlogP ≥ 8 [d)]	602 (2%)	275 (2%)	399 (4%)
Glycosides [e)]	8,260 (24%)	797 (5%)	1,793 (17%)
Peptides [f)]	194 (<1%)	676 (4%)	2,053 (19%)

a) COCONUT entries with a DOI and the specified taxonomical origin annotated; percentages refer to the total number of the selected entries within the specified class: 33,772 plants NPs, 15,648 fungal NPs, and 10,751 bacterial NPs. b) Molecular weight (MW) calculated with RDKit. c) Fraction of sp3 (Fsp3) calculated with RDKit. d) Octanol: water partition coefficient calculated with RDKit following the Crippen method (AlogP). e) Containing a cyclic N- or O-acetal substructure defined through SMARTS language. f) Containing a dipeptide substructure defined through SMARTS language.

To get a closer insight into structural features, we calculated the MAP4 fingerprint for each of the 60,171 selected NPs. MAP4 encoding combines the characteristics of substructure fingerprints, which are well suitable for small molecules, and of atom pair fingerprints, which are instead preferable for larger structures, and it has been proven suitable for both.[18] It consists of listing all pairs of circular substructures of radius 1 and 2 as SMILES, separated by their topological distance in bonds, and MinHashing the resulting set of SMILES pairs to a defined dimensionality (1024 in the present analysis). We then represented the MAP4 annotated NP dataset using the dimensionality reduction method TMAP. This method is suitable for very large high-dimensional datasets and performs better than t-SNE or UMAP in preserving local and global relationships in the data. [12] To create a TMAP, the algorithm computes an approximate nearest neighbor graph by locality sensitive hashing (LSH), cuts edges to obtain the minimum spanning tree of this graph, and creates an optimized 2D representation of the minimum spanning tree, in which each node represent a molecule connected to its approximate nearest neighbors. This tree is then displayed with interactive visualization tool Faerun. [24]

Faerun shows each node as a sphere that can be color-coded according to various properties and uses Smilesdrawer[99] to depict molecular structures. The TMAP of our NP subset is available interactively at https://tm.gdb.tools/map4/coconut_tmap/.

The TMAP of our NP subset color-coded by MW showed that most high MW compounds appeared in two groups, the first one (at right on the TMAP), contained peptides and related macrocycles, and the second one (at middle/lower left on the TMAP) corresponded to glycosylated triterpenoids (Figure 20a). Color-coding by Fsp3 showed that the TMAP separated high Fsp3 molecules (left half of the TMAP), comprising many terpenes, steroids, and glycosides, from low Fsp3 molecules (right half of the TMAP) featuring many polyphenols and related polyaromatic molecules (Figure 20b). Furthermore, the color-code by the calculated octanol:water partition coefficient AlogP, estimating polarity, showed several islands of highly polar NPs (low AlogP, magenta) corresponding mostly to nucleosides and glycosylated polyphenols (upper part of the TMAP), glycosylated triterpenoids (lower left on the TMAP) and peptides (middle right on the TMAP), as well as a few groups of apolar NPs (high AlogP, red), corresponding primarily to lipids, terpenes, and steroids (Figure 20c).

Color-coding by the annotated origin showed that NPs from plants, fungi, or bacteria formed many well-defined clusters spread across the entire TMAP (Figure 20d). On the one hand, this separation illustrated how NP origin corresponded to differences in molecular structure that were well perceived by the MAP4 fingerprint used to generate the map. On the other hand, the taxonomical origin color code also showed that each subset contained diverse structural types. While there was no correlation of origin with properties such as MW, Fsp3, or AlogP, most glycosides were associated with plants, and most peptides were of bacterial or fungal origin, in line with Table 9 (Figure 20e). These relationships were also well visible by color-coding the TMAP by prioritized categories (Figure 20f).

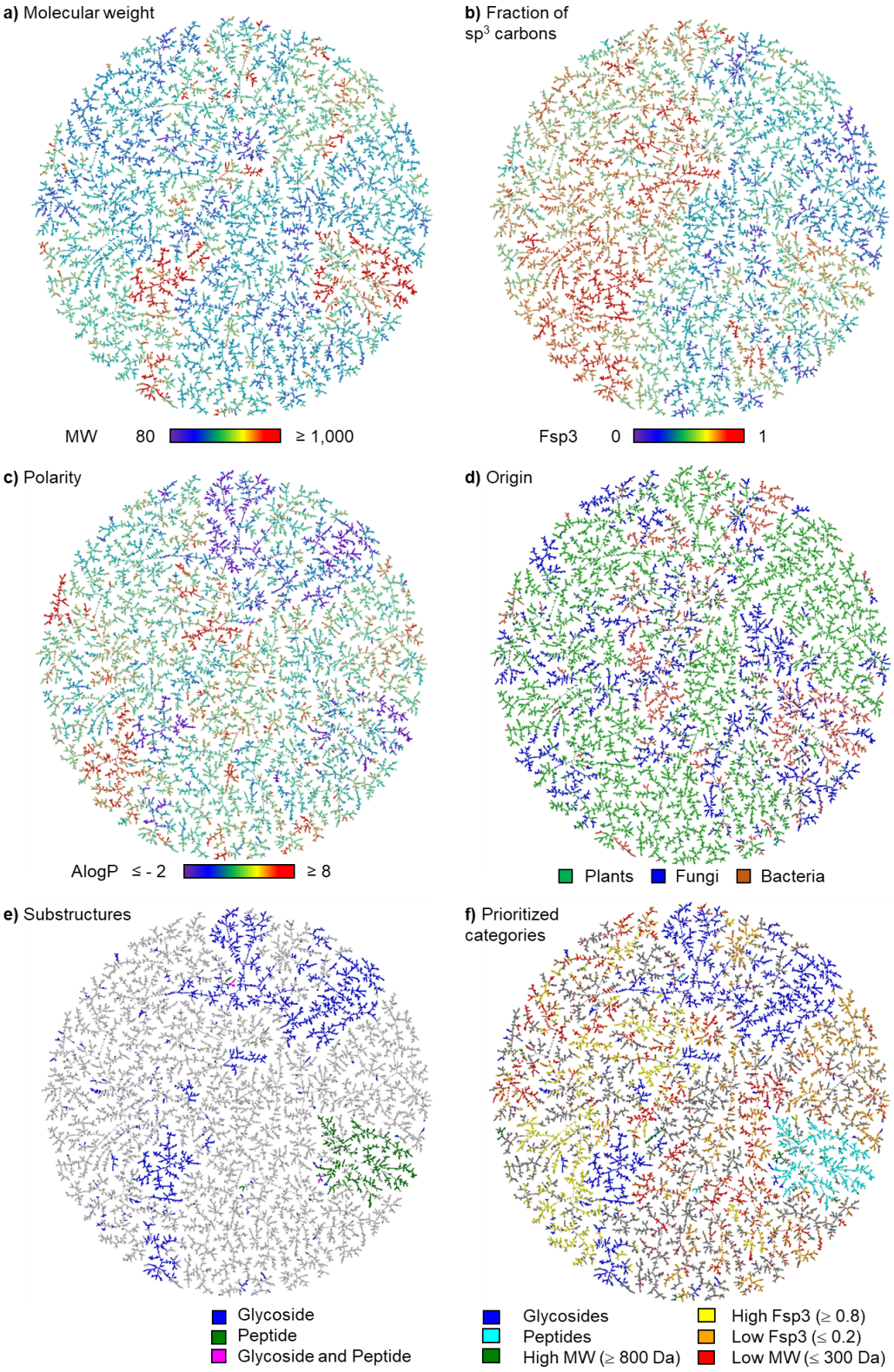


Figure 20. MAP4 TMAP of the 60 thousand selected COCONUT entries. The maps are colored according to (**a**) molecular weight MW in Da, (**b**) fraction of sp3 carbon atoms Fsp3, (**c**) calculated octanol:water partition coefficient AlogP, (**d**) COCONUT annotated origin, (**e**) presence of a glycoside (blue) or peptide (green) substructure, or both (magenta), (**f**) prioritized categories: glycosides (blue) > peptides (cyan) > high MW (green) > high Fsp3 (yellow) > low Fsp3 (orange) > low MW (red). Entries not belonging to any category are reported in gray. All maps are accessible in an interactive format at https://tm.gdb.tools/map4/coconut_tmap/.

5.2.2 Statistical modeling of NP origin using support vector machines (SVM)

The clear separation of NPs from plants, fungi, or bacteria in the TMAP above clearly showed that our MAP4 fingerprint distinguished between NPs of plant, bacterial or fungal origin. To further investigate this separation, we trained an SVM classifier using the MAP4 similarity matrix of half of the COCONUT subset and used the other half to evaluate it. Indeed, the obtained MAP4 SVM correctly predicted the origin of 94% of plant NPs, 89% of fungal NPs, and 89% of the bacterial NPs (MAP4 SVM), resulting in a balanced accuracy of 0.897, an MCC (Matthews correlation coefficient) of 0.890, and an F1 score of 0.920 (for a detailed explanation of the used metrics, please refer to the section 5.4.5).

To better identify the role of the MAP4 molecular encoding in the reported successful prediction, we compared the performances of a MAP4 SVM with the performances of an SVM trained using ECFP4 (Extended Connectivity Fingerprint ECFP of radius 2, ECFP4 SVM) and the RDKit atom pair fingerprint (AP SVM). We chose ECFP4 and the RDKit AP as widely used and available examples of substructures fingerprints and atom pair fingerprints. As a baseline model, we also included an SVM trained with a set of 11 calculated physico-chemical properties, namely MW, Fsp3, HBD (hydrogen bond donor) count, HBA (hydrogen bond acceptor) count, AlogP, the number of carbons, oxygens, and nitrogens, the total number of atoms, number of bonds, and TPSA (topological polar surface area) (properties SVM). The selected 60 thousand COCONUT entries were divided into five subsets, and each model was trained and evaluated five times using the five different 80-20 training test splits combinations of one subset as test set and the other four as training set. Then the mean balanced accuracy, MCC, and F1 score of the five evaluations were calculated.

The results of this evaluation are presented in Table 10 and Figure 21. Remarkably, all four SVM performed reasonably well. The good performance of the property based SVM

reflected the fact that relatively large NP families with characteristic properties are essentially all from the same origin. For example, almost all large peptides or cyclic peptides are assigned to bacteria, while most glycosylated triterpenoids and polyphenols are assigned to plants. Nevertheless, there was a significant performance increase with the ECFP4 SVM and MAP4 SVM, which performed best, showing that correct origin assignment works better if specific substructures are considered. Among the four SVM evaluated, our MAP4 SVM performed best with significantly higher values compared to the ECFP4 SVM, probably because the MAP4 fingerprint encodes a more precise representation of the molecular structures than ECFP4. Indeed, MAP4 considers pairs of local substructures and the topological distance between them, while ECFP4 only encodes the presence of local substructures.

Table 10. SVM evaluation with balanced accuracy, MCC, and F1 score.

	Balanced acc.	MCC	F1
MAP4 SVM [a,b]	**0.919**±0.005	**0.879**±0.005	**0.929**±0.003
ECFP4 SVM [a,b]	0.890±0.005	0.827±0.006	0.897±0.003
RDKit AP SVM [a,b]	0.735±0.005	0.592±0.006	0.752±0.004
Properties SVM [a,c]	0.758±0.005	0.613±0.007	0.761±0.004

[a] Mean value and standard deviation (σ) of the five different test/training sets split of the 5-fold cross-validation. [b] 1024 dimensions. [c] 11 properties: MW, Fsp3, HBD) and HBA, calculated logP with the Crippen method (AlogP), number of carbons, oxygen, and nitrogen, the total number of atoms, number of bonds, and topological polar surface area (TPSA).

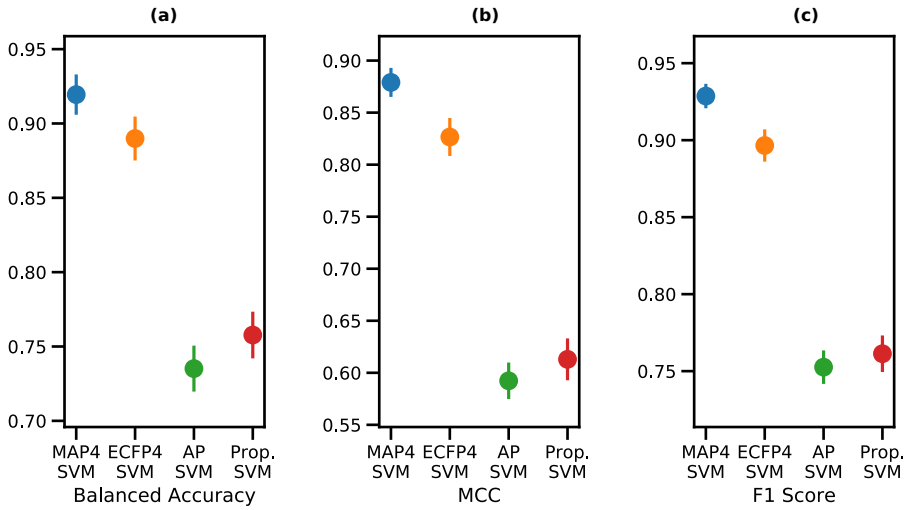


Figure 21. 5-fold cross-validation mean values and 3σ confidence intervals of the (a) balanced accuracy, (b) MCC, and (c) F1 score for the four SVM classifiers. In all panels, the MAP4 SVM is reported in blue, the ECFP4 SVM in orange, the RDKit AP (AP) SVM in green, and the properties (Prop.) SVM in red.

5.2.3 Using the MAP4 SVM to assign the origin of NPs

The SVM evaluation above showed that the MAP4 analysis of NP molecular structure identified features distinguishing between NPs assigned to plants, fungi, and bacteria. Assuming that most of the assigned origins were correct among the 60,171 NPs used for training, one may use an SVM to tentatively assign the origin of further NPs as originating from plants, fungi, or bacteria. To best exploit the information in the COCONUT database, we trained a MAP4 SVM using the entire set of 60 thousand COCONUT NPs assigned to plants, fungi, or bacteria. We used the resulting classifier to build an online tool that takes any molecular structure as input (drawn or pasted as SMILES) and returns the assigned origin and the corresponding percentages from the SVM classifier. This tool is freely accessible online at https://np-svm-map4.gdb.tools/.

The online tool performed quite well in assigning the origin of newly published NPs which were not present in COCONUT. Among thirteen recently reported NPs from plants, fungi, or bacteria, eleven were correctly assigned to their origin, while only two were misassigned (Table 11, Figure 22). In details, the fungal epicospirocin 1,[175] penicimeroterpenoid A,[176] and beetleane A,[208] the bacterial bosamycin A,[178] and the plant hunzeylanine A,[209] hyperfol B,[210] pegaharmol A,[211] mucroniferal A,[212] meloyunnanine A,[213] perovsfolin A,[214], and horienoid A[215] were correctly classified. On the other hand, the fungal rhizolutin[177] and the bacterial marinoterpin A[216] were misclassified as plant NP. Note that in these cases, the percentage values to the assigned class were lower than for the correct predictions.

Table 11. MAP4 SVM origin prediction for thirteen recently published microbial and plants NPs that are not present in COCONUT.

Natural Product	Origin	MAP SVM prediction[a)]
epicospirocin 1	fungal	fungal (97%)
penicimeroterpenoid A	fungal	fungal (82%)
beetleane A	fungal	fungal (97%)
rhizolutin	fungal	plant (55%, fungal: 29%)
bosamycin A	bacterial	bacterial (94%)
marinoterpin A	bacterial	plant (44%, bacterial: 37%)
meloyunnanine A	plant	plant (99%)
hyperfol B	plant	plant (93%)
pegaharmol A	plant	plant (77%)
mucroniferal A	plant	plant (73%)
hunzeylanine A	plant	plant (95%)
perovsfolin A	plant	plant (92%)
horienoid A	plant	plant (95%)

a) Predicted using the MAP4 SVM available online at https://np-svm-map4.gdb.tools/.

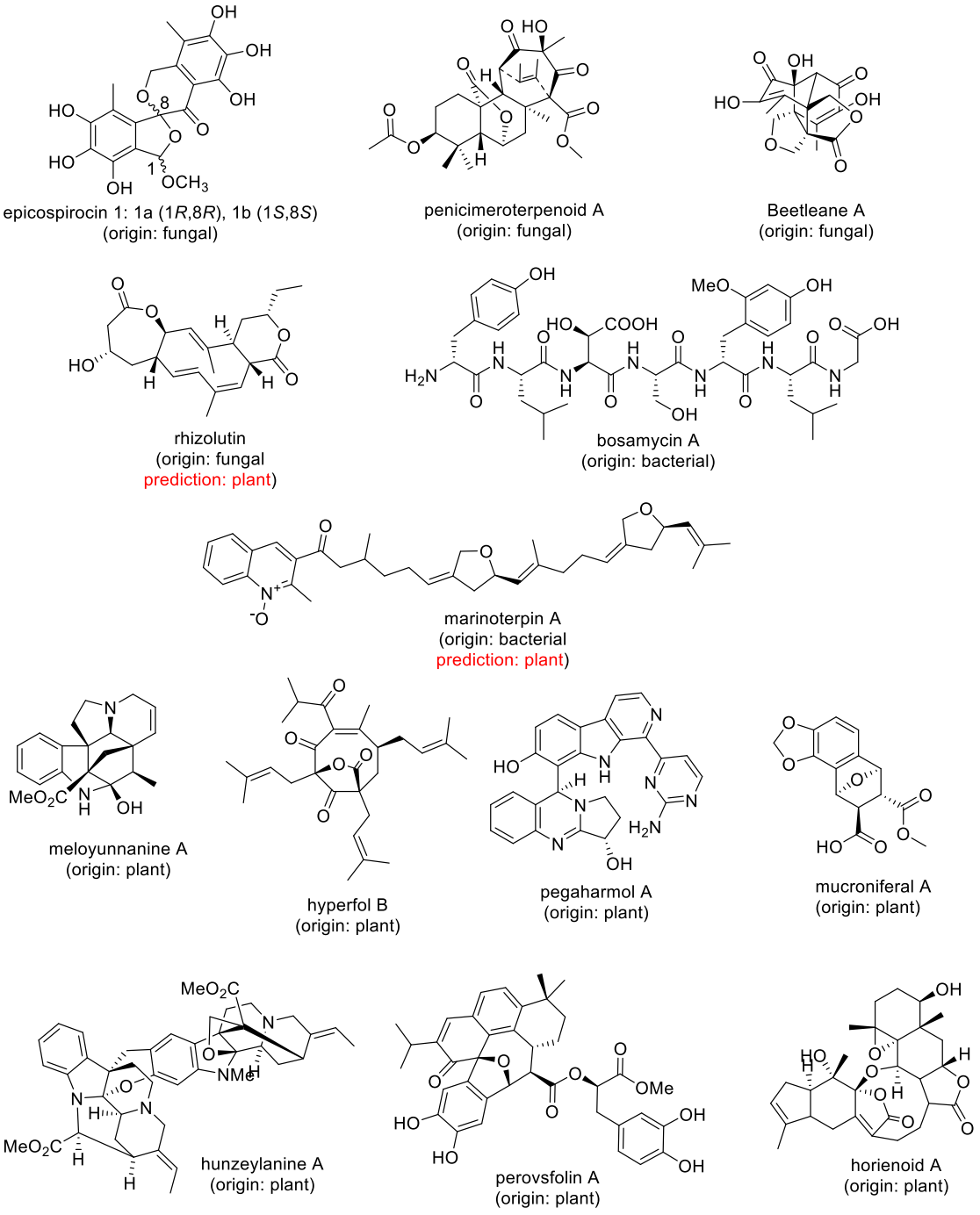


Figure 22. Chemical structure of thirteen recently published microbial and plants NPs which are not present in COCONUT. The MAP4 SVM prediction is identical with the origin unless marked otherwise.

In several cases, the SVM prediction conflicted with the superclass of the reported source organism. For example, the indole alkaloids cephalinones A-D and cephalandoles A-C isolated from the orchid *Cephalanceropsis gracilis*[217] and whose structures were partly revised by total synthesis,[218] were all assigned to bacteria by our SVM. In fact, These NPs might stem from an endophytic bacterium considering that endophytic microorganisms produce several related indole alkaloids.[219] Our SVM also reassigned the cancer drug maytansin from an annotated plant origin in the training set to a predicted bacterial origin, in line with its endophytic origin.[205] On the other hand, our classifier also assigned a bacterial origin to two cyclic peptides (CNP0085258 and CNP0085259)[220] and a cyclotide (CNP0085363)[221] isolated from plants. Although these plants indeed contain endophytic bacteria, the plant origin of such peptides is well established,[222,223] and the SVM assignment to bacteria reflects the fact that the majority of cyclic peptides and cyclotides in the COCONUT set used for training the SVM were assigned to bacteria, compared to only a handful of cyclotides of plant origin.

While the classifier may point to the possible endophytic origin of NPs isolated from plants, its use on NPs from other sources is problematic. For instance, among the 1,035 marine NPs from COCONUT with an annotated origin, 639 were assigned to plants by our SVM. This prediction must be mostly wrong considering that most marine organisms such as algae, corals, and sponges are not plants. For example, the 44 NPs from the soft coral *Sinularia*, or the macrocyclic terpene lactone lobophytolide A (CNP0275045) stemming from the soft corral *lobophytum cristagalli*,[224,225] were all incorrectly assigned to plants. However, the remaining 231 fungal and 165 bacterial predictions might be partly correct considering that many marine organisms carry endosymbionts. For example, our classifier assigned a bacterial origin for echinosulfonic acid B (CNP0318329), a brominated bis-indole NP isolated from the marine sponge *Echinodictyum gorgonoides*.[226] In this case, other authors have reported the isolation of a bacterial strain from the same sponge as a probable source of its biological activities.[227]

5.3 Conclusion

In summary, we visualized the chemical space covered by a subset of 60 thousand NPs from the COCONUT database with an assigned origin and publication using a TMAP calculated on the basis of MAP4 as molecular fingerprint, which is available at https://tm.gdb.tools/map4/coconut_tmap/. Analyzing this TMAP revealed that NPs from plant, fungal or bacterial origin form well separated groups. We then trained an SVM classifier with the MAP4 fingerprint to assign the origin of NPs and found that it performed excellently and significantly better than classifiers trained with ECFP4, RDkit AP, or physico-chemical descriptors.

To help assign NP origin, we then trained a MAP4 SVM classifier using the entire set of 60 thousand NPs. This tool is available online at https://np-svm-map4.gdb.tools/ and returns an origin prediction for any molecular structure drawn or pasted as SMILES. We found that this classifier correctly predicts the origin of plant, bacterial or fungal NPs not included in the 60 thousand COCONUT set used for training, as exemplified with the correct prediction of eleven out of thirteen newly published NPs. Broader testing of the classifier with further NPs from COCONUT showed limitations for NPs not from plant or microbial origin, such as marine NPs, but it also led to interesting use cases suggesting that the tool might serve as a help to assign NP origin. This concerns in particular NPs isolated from plant but which might in fact be produced by endophytic microorganisms. Such information could be essential when searching for the corresponding biosynthetic genes.

5.4 Methods

5.4.1 Database preprocessing

The coconut database was downloaded. Only the 135,091 (out of 400,837) entries having a taxonomical annotation were selected. The selected subset was further filtered down to the 67,730 entries having an annotation not shorter than ten characters in the DOI field. Then, the taxonomy field was split by commas and match towards the words "plant"/"plants", "fungi"/"aspergillus", "bacteria"/"bacillus"/"bacta" to select the NPs with an annotated plant, fungal, or bacterial origin, respectively. The entries common between multiple origins were assigned with the following priority: human > animal > bacteria > fungi > plant > marine. The process led to the selection of 33,772 plant NPs, 15,648 fungi NPs, and 10,751 bacterial NPs with annotated DOI, for a total of 60,171 structures. The number of carbons, oxygen, and nitrogens, the total number of atoms, number of bonds, and TPSA were extracted from the COCONUT annotations. MW, Fsp3, HBD, and HBA count, AlogP, were calculated using RDKit. [123] The presence/absence of a peptide or a glycoside moiety was evaluated using Daylight[158] SMILES arbitrary target specification (SMARTS) language. SMARTS were used with RDKit to identify COCONUT entries containing a dipeptide substructure, defined as "[NX3,NX4+][CH1,CH2][CX3](=[OX1])[NX3,NX4+][CH1,CH2][CX3](=[OX1])[O,N]", or a containing a glycoside defined as cyclic N- or O-acetal substructure with the SMARTS "[CR][OR][CHR]([OR0,NR0])[CR]".

5.4.2 Fingerprint calculation

The 1024 dimensions MinHashed atom pair fingerprint of radius 2 was calculated using the open-source code of MAP4.

5.4.3 TMAP layout

The indices generated by the MinHash procedure of the MAP4 calculation were used to create a locality-sensitive hashing (LSH) forest[42] of 32 trees. Then, for each structure, the 20 approximate nearest neighbors (NNs) in the MAP4 feature space were extracted from the LSH forest, and the tree layout was calculated. The LSH forest and the minimum spanning tree layout were calculated using the TMAP open-source code. Finally, Fearun[24] was used to display the obtained layout interactively.

5.4.4 MAP4 SVM implementation

The coconut SUBSET entries used to generate the TMAP were assigned to training or test set with a 50% random split. The SVM was trained using the MAP4 fingerprints of the training set. It utilized a custom kernel that calculates the similarity matrix between two MAP4 fingerprints, where the similarity of fingerprint *a* and fingerprint *b* is calculated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the number of elements of fingerprint *a*. The class weights were inversely proportional to the class frequency, and the hyperparameter C was optimized using 5-fold cross-validation. During the hyperparameter optimization, 20% of the training set was left out as a validation set, and the balanced accuracy of the validation set was maximized. The hyperparameter C was optimized among the values 0.1,1, 10, 100, and 1000, resulting in C = 1. The classifier was implemented using scikit-learn[162] with the "one versus rest" strategy, and all not mentioned hyperparameters were used in their default values. Platt scaling,[160] was used to obtain probabilistic prediction values. After the evaluation process, a second version of the MAP4 SVM classifier was trained using both training and test to learn from all curated 60 thousand data points.

5.4.5 MAP4, ECFP4, RDKit AP, and properties SVMs comparison

The MAP4, ECFP4, and the RDKit AP fingerprints and a set of 11 properties (MW, Fsp3, HBD and HBA count, AlogP, number of carbons, oxygens, and nitrogens, total number of atoms, number of bonds, and TPSA) were used to train four different SVM classifiers in a 5-fold cross-validation. For all classifiers, the class weights were inversely proportional to the class frequency, and the hyperparameters were optimized using 10% of the available data (Table 12). For the properties SVM, the 11 values were scaled to zero mean and unit variance.

Table 12. Non-default and optimized hyperparameters used in the 5-fold cross-validation MAP4, ECFP4, RDKit AP, and properties SVMs comparison.

Classifier	Kernel [a]	C [a]	γ [a]
MAP4 SVM	**MAP4** [b]	0.01, 0.1, **1**, 10, 100	-
ECFP4 SVM	**Tanimoto** [c], Dice [c]	0.01, 0.1, **1**, 10, 100	-
RDKit AP SVM	**Tanimoto** [c], Dice [c]	0.01, 0.1, 1, **10**, 100	-
Properties SVM	**RBF** [d]	0.01, 0.1, **1**, 10, 100	0.01, 0.1, **1**, 10, 100

a) Used hyperparameters are reported in bold. b) Calculates the similarity matrix between two MAP4 fingerprints, where the similarity of fingerprint *a* and fingerprint *b* is calculated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the number of elements of fingerprint *a*. c) Ralaivola *et al.* [228] d) Radial basis function (RBF) kernel.[161]

5.4.6 Classifiers evaluation metrics

The F1 score is defined as the harmonic mean of precision and recall:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

Where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives predicted by the classifier.

The balanced accuracy is defined as:

$$Balanced\ accuracy = \frac{\frac{TP}{TP + FP} + \frac{TN}{TN + FN}}{2}$$

The Matthews correlation coefficient (MCC) is a correlation between the observed and the predicted class and it is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5.4.7 Online MAP4 SVM

The MA4 SVM classifier trained with the whole 60 thousand COCONUT subset is found at https://np-svm-map4.gdb.tools/. The query molecule can be provided as a drawn structure or pasted SMILES in the JSME editor[132]. The given query is canonicalized, chirality information is removed with RDKit, and the MAP4 fingerprint is calculated. To obtain probabilistic prediction values for each class, we use Platt scaling.[160]

Part Two – Peptide Chemical Space and Generation of Novel Peptide Sequences

Review Chapter Six – Peptides in Chemical Space

This work is based on the peer-reviewed publication:

Capecchi, A.; Reymond, J.-L. Peptides in Chemical Space. Med. Drug Discov. 2021, 9, 100081. https://doi.org/10.1016/j.medidd.2021.100081

Abstract

Peptides and their chemical space are highly relevant in the field of medicine. The recent advances in computer hardware and software led to a wide application of computational methods for the analysis and the selection of bioactive peptides. In this review, we report different *in silico* strategies for the discovery of new bioactive peptides with a focus on structure-based design, genetic algorithms, machine learning, and the use of molecular fingerprints to sample virtual libraries. Then, we describe the chemical space of known peptides through an analysis of 11 available peptide and peptide-containing databases, and we provide an interactive MAP4 (MinHashed Atom Pair fingerprint of diameter 4) TMAP (Tree Map) of the over 40 thousand extracted unique sequences available at https://tm.gdb.tools/map4/peptide_databases_tmap/.

6.1 Introduction

Peptides, defined as sequences of amino acids up to approximately 50 residues in length, represent an extremely large reservoir of potentially bioactive compounds, referred to here as the peptide chemical space. The relevance of this chemical space for medicine is evidenced by a large number of therapeutic peptides, in particular hormones and analogs such as insulin[229] or the recently FDA-approved bremelanotide [230–232]. Despite its size, the peptide chemical space can be precisely defined through a list of amino acid building blocks, usually the proteinogenic amino acids, and the length and topologies of the peptide chains that are considered, which may be linear, cyclic, or branched.

Following the invention of solid-phase peptide synthesis (SPPS) and recombinant methods in molecular biology, a number of experimental approaches have been developed to search the peptide chemical space for compounds binding to a specific molecular target by synthesizing and testing large combinatorial libraries [233]. More recently, advances in computer hardware and software have made it possible to select bioactive peptides by computational methods [234–238], thereby focusing experimental evaluation to a selected set of test sequences, as well as to develop a global understanding of the peptide chemical space by comparing all known bioactive peptides with each other. In this review, we summarize recent advances in computational peptide design. We classified computational design approaches as follows: a) structure-based design, where 3D-modeling of the site of action guides the selection of test peptides; b) GA (genetic algorithms), which select test peptides by iterative cycles of mutations and fitness selection; c) ML (machine learning methods), which exploit information on known bioactive peptides to propose new ones; and d) molecular fingerprints, which focus on calculated molecular similarity between peptide structures to enable a focused sampling (Figure 23). Finally, we present an overview of the currently know chemical space of bioactive peptides in form of an interactive chemical space map.

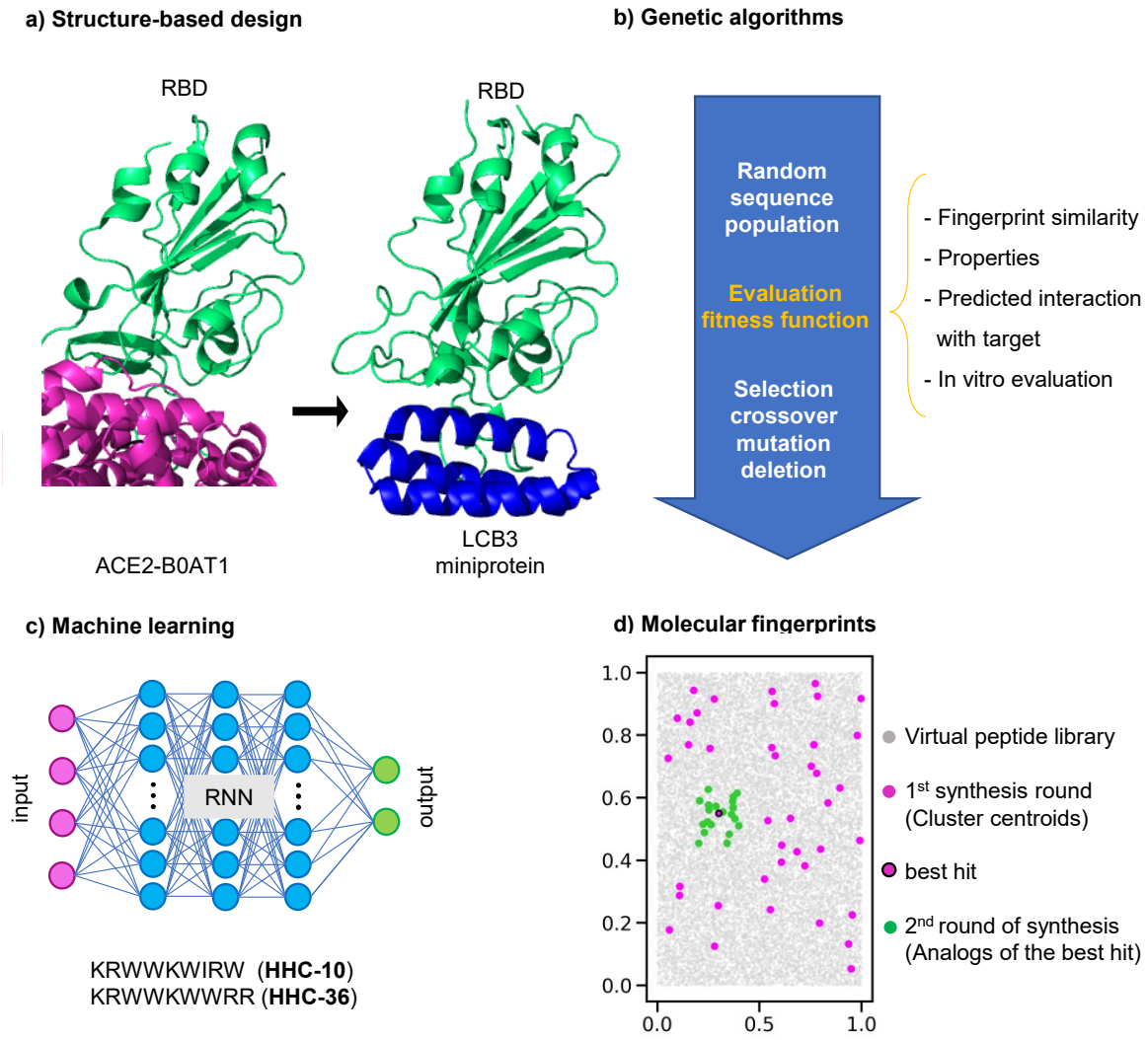


Figure 23. (**a**) The SARS-CoV-2 spike protein RBD (Receptor Binding Domain, green) binds ACE2 (Angiotensin-converting enzyme 2, pink), causing the virus to enter the cell. The miniprotein LCB3 (blue) was designed to bind the RBD (green) and inhibit the interaction between RBD and ACE2. (**b**) Genetic algorithm workflow schematic and classification based on the fitness function. (**c**) Example of an RNN (Recurrent Neural Network) architecture and sequences of the antimicrobial HHC-10 and HHC-36 discovered with this approach. (**d**) schematic representation of the sampling of a virtual library using molecular fingerprints.

6.2 Structure-Based Design

If the 3D-structure of the targeted site of action of the desired peptide is known in advance, one can select potentially bioactive peptides by modeling their interactions with this site using docking and molecular dynamics. This approach has been historically the first method to design

bioactive peptides computationally and has been extensively reviewed [239–243]. A recent example of this approach is the computational design by Cao et al. of miniprotein inhibitors of the SARS-CoV-2 spike protein ACE2 (Angiotensin-converting enzyme 2) interaction stopping the viral entry into cells [244]. The inhibitors were designed *in silico* using two different strategies. Firstly, a library of peptide sequences was designed using the Rosetta software [245] to incorporate the ACE2 helix responsible for the most interaction with the spike protein RDB (Receptor Binding Domain). Secondly, a set of sequences was designed from scratch though large large-scale de novo design of small helical scaffolds, followed by RIF (Rotamer Interaction Field) docking with the spike protein RBD (receptor binding domain), where RIF docking has the peculiarity of considering multiple conformations of the binding pocket (Figure 23a) [246].

Other recent examples of structure-based peptide design include the discovery of cyclic peptides with high binding affinity to diverse influenza strains through modeling based on antibody loops by Sevy *et al.* [242], and the design of stapled peptides that activate the VapC complex of the *Mycobacterium Tuberculosi* and lead to the arrest of bacterial cell growth by Kang *et al.* [247]. Structure-based design sometimes simply aims to identify peptides that mimic the structure of a known bioactive peptide. A recent example of this approach is the design of peptides that assemble into cross-α amyloid-like structures by Zhang *et al.* [248].

6.3 Genetic Algorithms

A GA is a search algorithm inspired by the evolution theory which optimizes a population of solutions towards a given goal through iterative cycles of mutations and selection of the fittest solutions using a fitness function [249]. If the solutions searched by the algorithm are set to be peptide sequences, GAs can be applied to find novel peptides (Figure 23b).

The fitness function of a peptide GA can be based on calculated properties. In 2003, Teixido *et. al.* used a GA to identify peptides capable of crossing the BBB (blood-brain barrier) with a fitness function based on a set of descriptors comprising molecular weight, length, amphiphilicity, isoelectric point, LogP, secondary structure, presence of aromatic and positive residues, potential hydrogen bonds, and the nature of C- and N-termini. The ideal set of values for these descriptors was derived from a statistical analysis of the experimental data on peptide-BBB permeability [250]. More recently, Beltran and Brizuela used mean hydrophobicity, helical hydrophobic moment, net charge, and isoelectric point to design selective cationic antibacterial peptides [251]. In another recent example of GA guided by properties, Port *et al.* optimized a guava antimicrobial peptide using a fitness function based on the ratio between hydrophobic moment and α-helix propensity [252].

Predicted protein-peptide interaction can also be used as fitness function of a GA. In 2011, Knapp *et al.* optimized peptides for major histocompatibility complex binding using a GA and the consensus of five different binding prediction methods in its fitness function [253]. More recently, King *et al.* discovered an α-conotoxin analog with optimal binding to the α3β2-nicotinic acetylcholine receptor using a GA and an AutoDock based fitness function to guide their search [254].

The fitness function used to guide selection in a GA can also be estimated with ML (machine learning) property prediction. For example, Fjell *et al.* used the prediction of an artificial neural network to drive a GA towards active antimicrobial peptides [255]. Additional ML approaches are discussed in the following sections. A further example of GA for peptide design exploiting molecular fingerprint similarity as fitness function is discussed below in section 6.5.

It is also possible to use the feedback of experimental analysis to guide a GA as exemplified by the work of Yoshida *et al.* that combines supervised ML with in vitro testing as fitness function of GA to optimize antimicrobial peptides [256]. Another recent example is the work of Neuhaus *et al.* Starting from known ACPs (anticancer peptides), the authors used a GA coupled with in vitro testing to generate new ACPs with improved activity, and they showed that both the interaction with the membrane and peptide dimerization degree were responsible for the anticancer activity [257].

6.4 Machine learning

ML approaches are used for two major tasks: property prediction and generation of new sequences. For the first case of property prediction, one uses supervised ML techniques, for which the task consists of mapping an input to a specific output (Figure 23c) [258–263]. The input can be the peptide sequence itself, but also descriptors, structure-based features, molecular fingerprints, or a combination of the previous. The output of the ML model is usually a label, such as active/inactive for a specific application. Property prediction by ML requires a large amount of highly curated data, highlighting the importance of manually curated peptide databases that collect sequence activity and toxicity.

The first example of this approach was reported in 2009 by Cherkasov *et al*. with the discovery of two tryptophan- and arginine-rich antimicrobial peptides, HHC-10 and HHC-36, which were more potent and shorter than similar arginine-rich peptides found in Nature such as indolicidin (Figure 23c) [264]. The authors trained an artificial neural network classifier with 44 QSAR descriptors to discern between antimicrobial and non-antimicrobial peptides, and then they used this trained neural network to classify each peptide in a virtual library of 100,000 random nonapeptides enriched with tryptophan, arginine, and lysine, as active or inactive. In a

recent example of this approach [265], Timmons and Hewage showed that one can use supervised ML to train a neural network classifier to distinguish between hemolytic and non-hemolytic peptides at the example of peptides from the DAASPDB and the Hemolytic [266] databases.

The second application of ML consists of training generative models to output new peptide sequences with specific characteristics. For example, Müller *et al*. recently reported an LSTM-RNN (long short-term memory recurrent neural network) capable of generating helical peptides with predicted antimicrobial activity [267]. In a similar approach, Grisoni *et al*. trained an LSTM-RNN to generate alpha-helical cationic amphipathic sequences, and then they fine-tuned it using 26 known ACPs. Twelve of the proposed sequences were synthesized and ten showed the expected membranolytic activity [143].

Classification and generative ML models can also be combined. For example, Tucs *et al*. recently reported a GAN (Generative Adversarial Network) to generate antimicrobial peptides [268]. A GAN is a ML architecture composed of a ML generative model and a discriminator, which is generally a ML classifier, whereby both models are trained as a pair. The task of the generator is to generate sequences resembling known antimicrobial peptide sequences, while the task of the discriminator is to distinguish between potential antimicrobial peptides and random sequences.

6.5 Molecular fingerprints

We recently showed that one can discover bioactive peptides computationally in the absence of precise structural modeling by using molecular fingerprint comparisons. This approach is well-known in small molecule drug discovery [117] but still underexploited with peptides. We demonstrated the feasibility of this approach by discovering antimicrobial bicyclic peptides against the Gram-negative bacterium *Pseudomonas aeruginosa* and its biofilms[6].

To discover active bicyclic peptides, we used a shape and pharmacophore fingerprint called 2DP describing the relative positions of cationic and hydrophobic groups, an important parameter for the targeted membrane disruptive activity. The workflow comprised the following steps: 1) establishing a SPPS protocol for bicyclic peptides comprising nine variable positions; 2) enumerating a virtual library considering all possible combinations of lysine, leucine at the variable positions; 3) computing 2DP-fingerprint similarities between all pairs of bicyclic peptides and clustering the virtual library to sample the overall diversity of the virtual library; 4) synthesizing and testing a small set of sampled bicyclic peptides. This approach led to the identification of a single active bicyclic peptide, which we then optimized by synthesizing and testing further analogs identified by 2DP-similarity searching in the virtual library (Figure 23d). The virtual library of this proof-of-concept experiment only comprised 6,230 different bicyclic peptides. In a subsequent project, we applied the same approach to a differently designed and much larger virtual library of 4.7 million bicyclic peptides and identified the cysteine bridged bicyclic peptide **bp50** and its D-enantiomer **bp56** as a potent antimicrobial peptide against multidrug-resistant strains of *A. baumannii* and *P. aeruginosa* (Figure 24a) [5]. Note that membrane disruptive peptides are generally conformationally flexible and that this conformational flexibility is necessary for their activity [269]. Indeed, molecular dynamics studies and CD-spectra suggest that **bp56** exists in a dynamic equilibrium between a β-sheet conformation in water and a partially α-helical amphiphilic conformation in a membrane environment (Figure 24b).

Considering that our 2DP molecular fingerprint could be applied to any type of peptide chain topology, we further implemented this fingerprint-based approach to search for analogs of AMPD (antimicrobial peptide dendrimer) **G3KL**, which contains a highly ramified peptide chain [270]. This AMPD kills a broad range of Gram-negative bacteria including multidrug-resistant clinical isolates by a membrane disruptive mechanism with almost no resistance

[3,271,272], and it exhibits angiogenic as well as antibiofilm properties [273,274]. By generating a virtual library of **G3KL** analogs and testing 2DP-nearest neighbors of **G3KL**, we identified AMPD **T7** exhibiting enhanced serum stability and a broader activity spectrum (Figure 24c)[4]. Interestingly, AMPD **T7** corresponds to minor sequence changes at the dendrimer core compared to **G3KL**, which were thought to be negligible by design but turned out to have a major impact on antimicrobial activity.

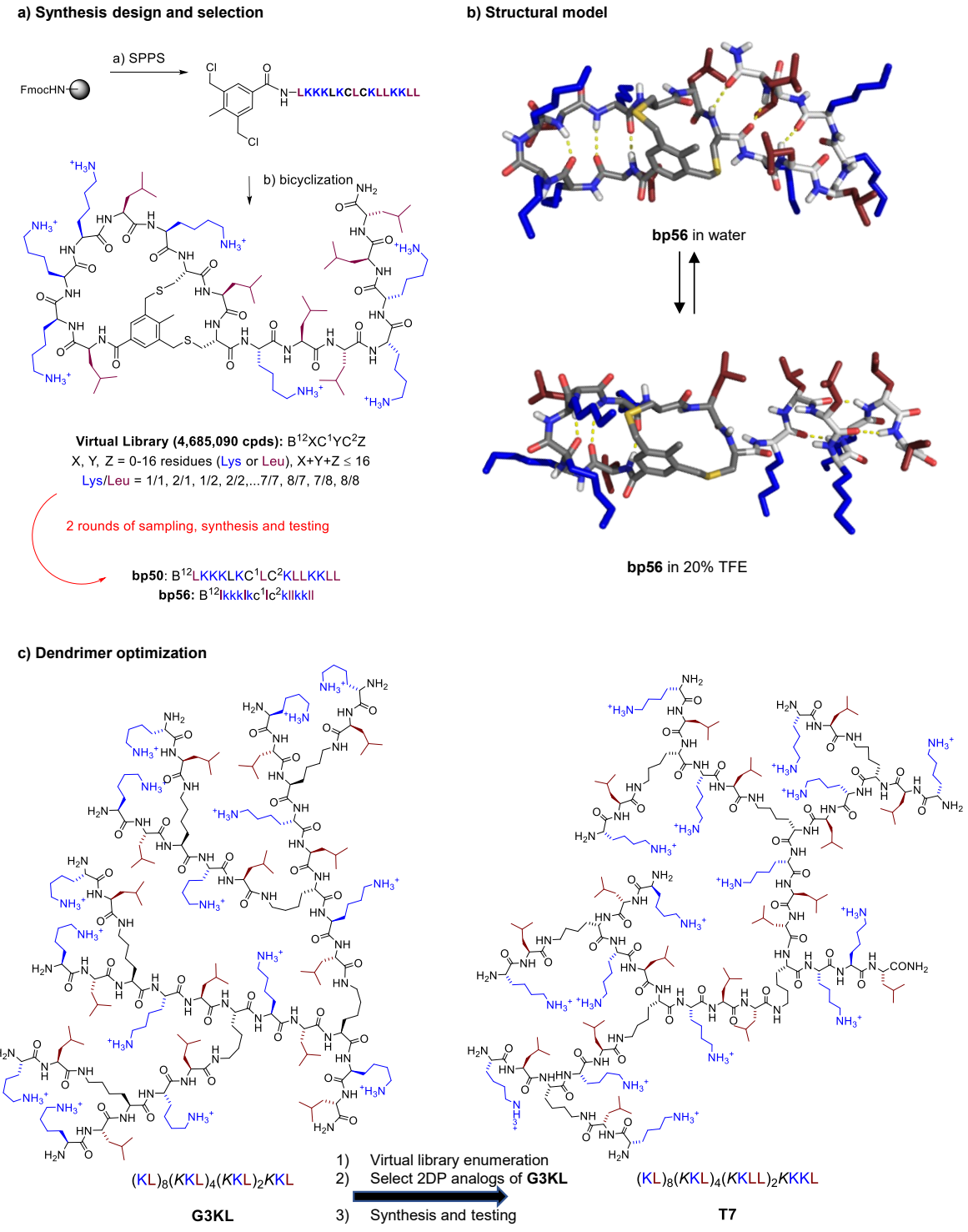


Figure 24. Molecular fingerprint guided discovery of antimicrobial peptides. (**a**) Synthesis and virtual library design and selection of **bp56**. (**b**) Molecular dynamics studies of **bp56** in water with or without TFE (trifluoroethanol) to mimic the membrane environment reveals a dynamic conformation. (**c**) Optimization of antimicrobial peptide dendrimer **G3KL** by virtual library enumeration, nearest neighbor selection, synthesis, and testing.

We recently implemented molecular fingerprint similarity as a fitness function in a GA called PDGA (peptide design genetic algorithm) capable of generating peptides of diverse topologies (linear, cyclic, polycyclic, or dendritic) resembling any target molecule of choice. In a typical PDGA run, all generated sequences above a defined molecular fingerprint similarity threshold are identified as analogs. Because the molecular fingerprint can be computed for any molecule of interest, PDGA can generate peptide analogs of both peptides and non-peptides[25]. Furthermore, PDGA operates with diverse peptide topologies including linear, cyclic, or polycyclic peptides as well as peptide dendrimers. In a proof-of-principle computation, we showed that PDGA generates known analogs of the cyclic peptide tyrocidine A and peptide dendrimer **G3KL**.

6.6 Visualizing the peptide chemical space

The ability to compute similarities between peptides allows representing the peptide chemical space in the form of maps in which distances represent similarities. Such maps provide an overview that helps to perceive the structural diversity of peptides. In our first implementation of this approach, we created an interactive map of the Protein Data Bank chemical space based on computed 3D-shape similarities [98]. However, this representation was only applicable to macromolecules with known 3D-structures such as those in the Protein Data Bank.

To represent peptide structural diversity in a general context, we have used the molecular shape similarity fingerprint used above with PDGA to compute similarities between molecules featured in non-Lipinski part of the ChEMBL and PubChem databases, which comprise 376,504 respectively 15,798,352 entries, 16 % respectively 7 % of which contain a dipeptide substructure [14]. These similarity comparisons can be represented in interactive 3D-

maps displayed using Faerun[24], in which each molecule appears as a point color-coded by a property of choice, and its structure is displayed using Smilesdrawer [99].

More recently we created a high-resolution molecular fingerprint called MAP4 useful to analyze diverse molecular classes spanning from small molecule drugs to metabolites, natural products, and macromolecules including peptides, DNA, and oligosaccharides [18]. The MAP4 fingerprint can be used in combination with the TMAP mapping tool [12] to create insightful representations of molecular databases, as recently shown for the case of the Natural Product Atlas [21].

For the present review, we have collected bioactive peptides from eleven publicly accessible databases that cover a wide range of size and scope (Table 13) [275,130,276–284]. We considered 40,531 database entries corresponding to sequences of between 2 and 50 natural amino acids, calculated their SMILES representation using RDKit [123], and used this data to compute a TMAP based on the MAP4 fingerprint. This organizes peptides by their size and sequence (Figure 25a, https://tm.gdb.tools/map4/peptide_databases_tmap/). The map colored by source database shows that several databases tend to cover specific regions of the peptide chemical space, which is not surprising for activity-specific databases such as SPdb in which sequences have limited diversity, but somewhat surprising for the peptides retrieved from PDB (Figure 25b). Color-coding the map by the number of databases in which a peptide is listed shows that most peptides (60%) occur only in one database, while 11% are present in two databases, 3.4% in three, 1.4% in four, and less than 1% in five databases (Figure 25c). Color-coding by activity type illustrates that the largest fraction of peptides in these databases (17,260 sequences, 43 % of the total) are annotated as antimicrobial and anticancer, and stem from the DBAASP, DRAMP, AVPdb, and the antimicrobial and anticancer sections of the SATPdb (Figure 25d and online map).

Table 13. Peptide and peptide-containing databases publicly available and downloadable in bulk.

Name	Description	Size[a)]	Web Page	Ref.
PDB[b)]	3-D structural data of large biomolecules	8,805	https://www.rcsb.org/	275
SwissProt[c)]	Sequences and functional information of peptides and proteins m**anually annotated**	9,129	https://www.uniprot.org/	130
SATPdb[d)]	Therapeutic peptides	14,985	http://crdd.osdd.net/raghava/satpdb/	276
DBAASP[e)]	Antimicrobial peptides	10,999	https://dbaasp.org/	277
DRAMP[f)]	Antimicrobial peptides	3,673	cpu-bioinfor.org	278
AVPdb[g)]	Antiviral peptides	1,801	http://crdd.osdd.net/servers/avpdb/	279
SPdb[h)]	Signal peptides	2,340	http://proline.bic.nus.edu.sg/spdb/	280
NeuroPedia[i)]	Neuropeptides	392	http://proteomics.ucsd.edu/Software/NeuroPedia/	281
DADP[j)]	Anuran defense peptides	743	http://split4.pmfst.hr/dadp/	282
Quorumpeps[k)]	Quorum sensing peptides	243	http://quorumpeps.ugent.be/	283
AntiAngioPred[l)]	Angiogenic peptides	197	http://clri.res.in/subramanian/tools/antiangiopred/index.html	284
Total of peptidic entries constituted by 2 to 50 natural amino acids		53,307		
Unique sequences collected across databases		40,531		

[a)] Number of unique peptidic entries constituted by 2 to 50 natural amino acids. [b)] PBD = Protein Data Bank. [c)] SwissProt = peptide sequences from the Uni-Prot database. [d)] SATPdb = Structurally Annotated Therapeutic Peptides database. [e)] DBAASP = Database of Antimicrobial Activity and Structure of Peptides. [f)] DRAMP = Data Repository of Antimicrobial Peptides. [g)] AVPdb = Antiviral Peptide database. [h)] SPdb = Signal Peptide database. [i)] NeuroPedia = Neuropeptides database and spectral library. [j)] DADP = Database of Anuran Defense Peptides. [k)] Quorumpeps = Quorumpeps database. [l)] AntiAngioPred = Server for Prediction of Anti-Angiogenic Peptides.

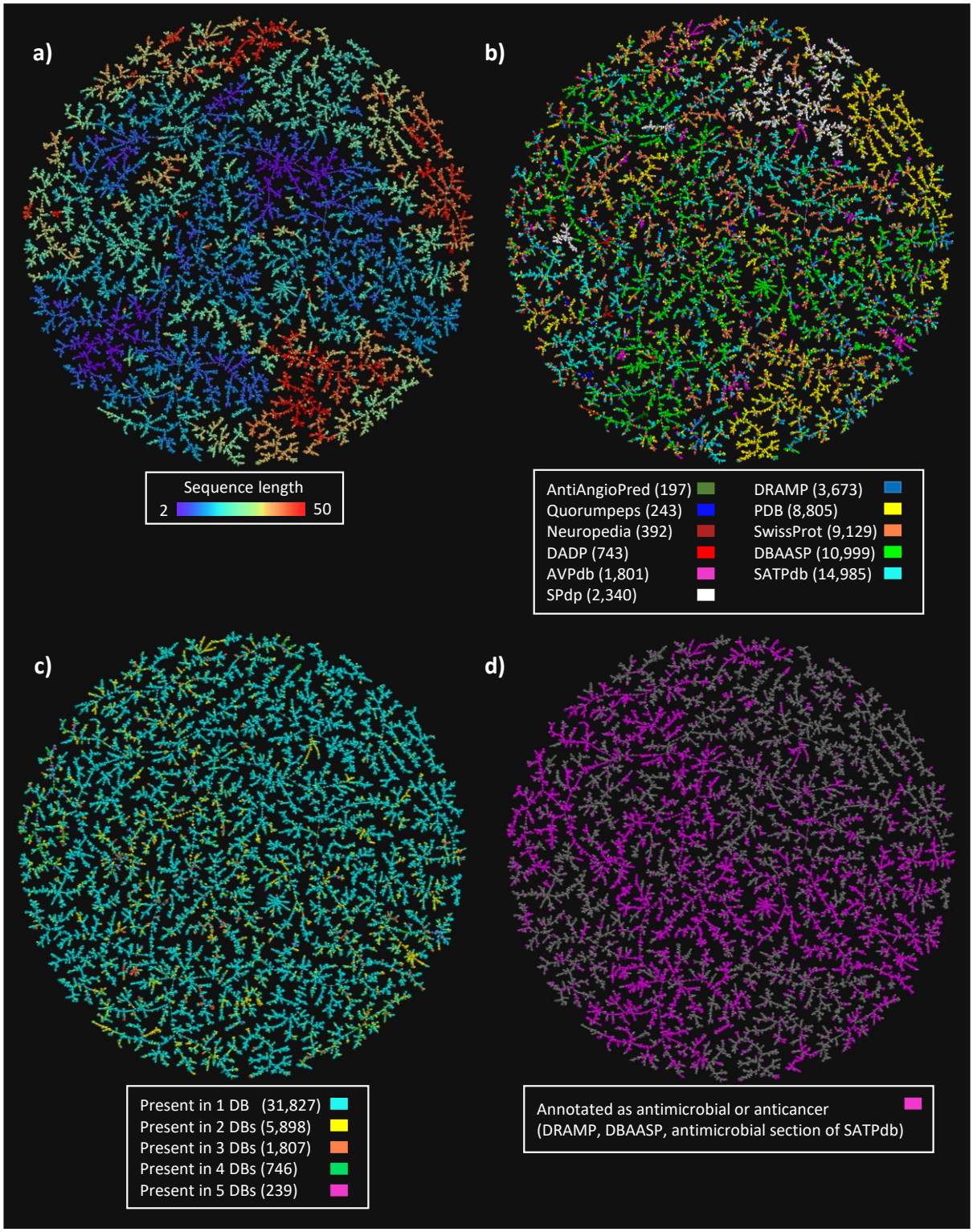


Figure 25. TMAP of the MAP4 encoded peptides databases space colored according to (a) sequence length, (b) Source database (DB, entries present in multiple databases were assigned to the smallest one), (c) occurrences across databases, (d) antimicrobial and anticancer activity. Further colors based on different activity criteria are available, and they can be found in the TMAP at https://tm.gdb.tools/map4/peptide_databases_tmap/.

6.7 Conclusive Remarks and Future Perspectives

In this review, we presented computational approaches to explore the peptide chemical space. Structure-based designs are well-suited when detailed information exists on the targeted site of action. GAs on the other hand have broader applicability since they can be used to design peptide sequences even if the targeted activity is not defined by a structure but more generally by a set of properties. ML is similarly broad in its applicability but requires a large number of known peptides with documented activity to enable model training. Finally, molecular fingerprints can be used to guide the sampling of large virtual peptide libraries as well as the optimization of known actives, as well as to compute graphical representations in the form of a map that facilitate a global understanding of the peptide chemical space. Most interestingly, GAs, ML, and molecular fingerprint-based approaches are possible without detailed knowledge of the peptide 3D-structure and allow to explore diverse peptide chain topologies, also incorporating non-natural amino acids. Such computational methods can play an enabling role in expanding the reach of peptides for therapeutic applications.

Chapter Seven – Populating chemical space with peptides using a genetic algorithm

This work is based on the peer-reviewed publication.

Abstract

In drug discovery one uses chemical space as a concept to organize molecules according to their structures and properties. One often would like to generate new possible molecules at a specific location in chemical space marked by a molecule of interest. Herein we report the peptide design genetic algorithm (PDGA, code available at https://github.com/reymond-group/PeptideDesignGA), a computational tool capable of producing peptide sequences of various topologies (linear, cyclic/polycyclic or dendritic) in proximity of any molecule of interest in a chemical space defined by MXFP, an atom-pair fingerprint describing molecular shape and pharmacophores. We show that PDGA generates high similarity analog of bioactive peptides with diverse peptide chain topologies, as well as of non-peptide target molecules. We illustrate the chemical space accessible by PDGA with an interactive 3D-map of the MXFP

property space available at http://faerun.gdb.tools/. PDGA should be generally useful to generate peptides at any location in chemical space.

7.1 Introduction

In drug discovery chemical space represents the ensemble of all molecules of possible interest as drugs.[285,286] The structural diversity of drugs and therefore their chemical space is extremely large and potentially overwhelming.[45,287] One possibility to gain an overview of chemical space is to artificially reduce its complexity by focusing on a subset of molecular properties which one can use to construct a mathematical "property space" which is also called "chemical space". Such spaces are most often high dimensional because one uses multiple properties to describe molecules. Nevertheless, dimensionality reduction methods enable to represent these spaces as 2D or 3D-maps and lead to a geographical understanding of molecular diversity because molecules that are found close to one another have similar properties.[288,103,289,46,118,290,24,12,105] The pertinence of this approach is supported by the fact that simple nearest neighbor searches in chemical spaces defined by high dimensional molecular fingerprints often perform as well or even better in virtual screening and target prediction benchmarks than more complex machine learning algorithms.[7,16,90,291,8,120]

Given a compound of interest, one would often like to generate new molecules at the same location in chemical space. In the area of small molecules one can identify such close analogs by virtual screening of possible molecules listed in computational combinatorial libraries[287,292] or generated on demand using deep neural networks.[293–295,149] The same approaches can be in principle applied to larger molecules beyond Lipinski's rule of 5 limit[13] considered in the present report, which are of interest as new modalities to address targets that are not druggable with small molecules.[71,296,297] However, the number of possible molecules and their structural diversity increases exponentially as function of molecule size.[298] Therefore, for large molecules one must first focus on well-defined subsets defined by a family of building blocks and the corresponding coupling chemistry before considering a computational strategy.

Such focus limits structural diversity but at the same time ensures that the proposed molecules should be synthetically accessible.

The most common subset of large molecules is that of peptides consisting of chains of amino acids linked by amide bonds. Machine learning, genetic algorithms and artificial intelligence have been used previously to design new peptides sequences with targeted structural and functional properties, mostly antimicrobial and anticancer activities. However, the reported examples were entirely focused on short linear peptides and often required large training sets of active compounds to produce new sequences.[264,252,299,300,256,301–304,236] Herein, we report a genetic algorithm capable of generating high similarity peptide analogs of diverse large molecules, including not only linear peptides but also cyclic/polycyclic peptides and peptide dendrimers and even non-peptides. Our peptide design genetic algorithm (PDGA) requires only a single target molecule as input.

PDGA generates peptides with diverse topologies of the peptide chain (linear, cyclic, polycyclic or dendritic) with high similarity to this target by using as fitness function a similarity calculated using MXFP (macromolecule extended atom-pair fingerprint). MXFP is a molecular fingerprint which uses the principle of atom pairs and is suitable for mapping the chemical space of large molecules beyond Lipinski's rule of five limit in ChEMBL and PubChem. [14] Atom pair fingerprints consider pairs of atoms in a molecule and the topological distance separating them counted in bonds. Despite of the fact that only topological 2D-information is considered, atom-pair fingerprints have the ability to represent 3D-molecular shape and pharmacophores, which often correlate with biological activities.[10,121,305,306,9,97,98,307] Most importantly, we recently used atom-pair fingerprint similarity searching to discover and optimize antimicrobial bicyclic peptides and peptide dendrimers against multidrug resistant Gram-negative bacteria despite the fact that these peptides do not have a well-defined folded conformation and that their activity in fact depends on conformational flexibility.[4–6] These

experiments provided an important validation of atom-pair fingerprint guided searches for practical peptide discovery even in the absence of well-defined 3D-conformations. In the present report we show that PDGA, guided by MXFP similarity as fitness function, provides the means to populate the non-Lipinski chemical space with peptides of diverse topologies in a targeted manner.

7.2 Results and Discussion

7.2.1 Peptide design genetic algorithm (PDGA)

Our Peptide Design Genetic Algorithm (PDGA, available at https://github.com/reymond-group/PeptideDesignGA) starts with a target molecule and a population of random peptide sequences with user-defined topology of the peptide chain (linear, cyclic, or dendritic), and performs rounds of modification and MXFP-similarity selection until the target or a preset time limit has been reached. The algorithm performs amino acids point mutations, insertions, deletions, cross-over, and cyclization/linearization or insertion of a branching unit depending on the selected topology, on character strings representing peptides. In these strings each character represents a building block. PDGA can use any natural or non-natural amino acid (e.g. β-, γ-, ω-amino acids, any non-natural side-chain, see Table 24 for the selection of building blocks used in this study), and includes variations in topology of the peptide chain by considering C to N cyclization and bridging cysteines for cyclic and polycyclic peptides and branching diamino acids (such as lysine) for dendrimers. Note that chirality is not encoded by MXFP and therefore not included in the PDGA output.

In its main implementation PDGA uses MXFP similarity to the target molecule as fitness function. To calculate MXFP similarity PDGA performs the following operations: 1) convert the character string to a SMILES taking topology of the peptide chain into account; 2)

assign atomic properties (H-bond donor, H-bond acceptor, positive and negative charges, aromaticity, and hydrophobicity) to each atom by assigning substructures to precomputed residues using SMARTS and compute MXFP values considering the properties assigned to each atom; 3) calculate the Manhattan distance (city-block distance, CBD) to the target molecule.

In all case studies described herein, we found that PDGA operates best with a population of 50 sequences with the same topology as the target molecule, retaining the 10 fittest sequences for the next round, inserting 5 new random sequences per generation, and generating the rest of the new population through mutation and crossover. PDGA stops after 24 hours if the target has not been reached. Sequences generated with a distance value $CBD_{MXFP} \leq 300$ to the target generally correspond to interesting analogs, which are stored to constitute the analog database. All studies presented below use this set of value for these PDGA parameters (Figure 26).

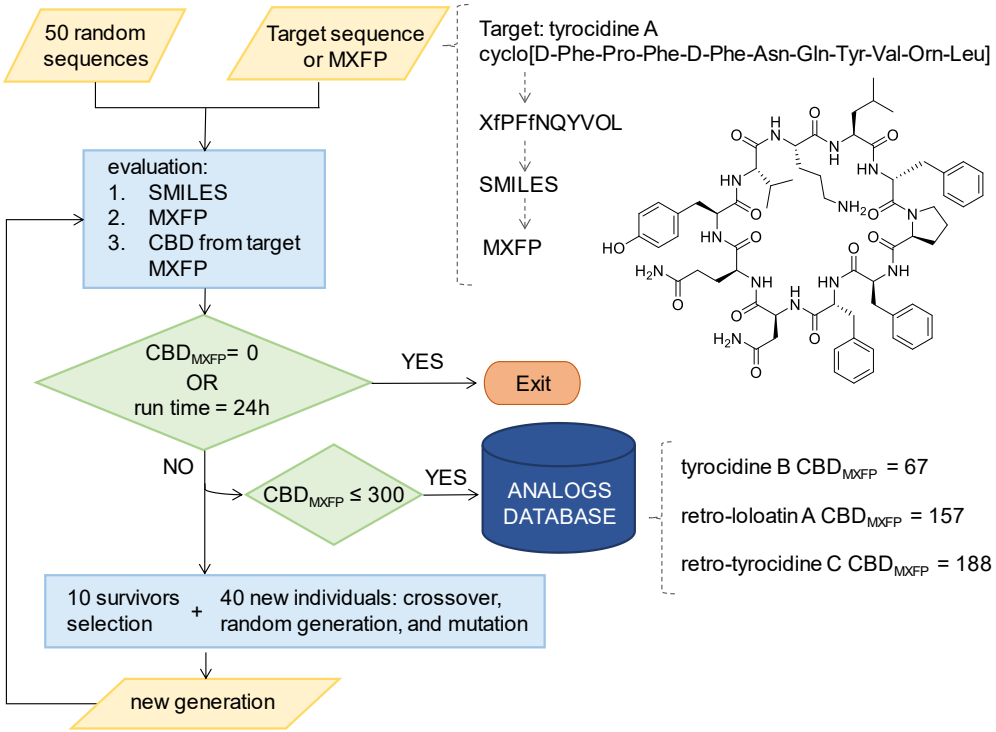


Figure 26. PDGA flowchart and analogs generation. 24 hours, 50, 10, 40, and 300, are variables depending on input parameters and settings. The input/output of the algorithm is exemplified for the target structure tyrocidine A.

We exemplify PDGA with four peptides of different topologies as targets, namely linear peptide indolicidin,[308] cyclic peptide tyrocidine A,[309] polycyclic peptide ω-conotoxin-MVIIA,[310] and peptide dendrimer **G3KL**, running the algorithm with the corresponding PDGA topology subclass (Table 14).[270] In all cases the minimum CBD_{MXFP} per generation decreases with increasing generation number, reflecting the progress of the genetic algorithm driven optimization (Figure 27a). Remarkably, PDGA mostly generates peptides within the $CBD_{MXFP} \leq 300$ limit, which represents significant target similarity. Indeed, by comparison randomly generated peptides have an average distance of $CBD_{MXFP} \sim 1000$ to the targets and almost no occurrence within the $CBD_{MXFP} \leq 300$ limit (Figure 27b-d). Note that most runs do not reach their target before the 24 h limit, nevertheless these runs are useful because they produce large numbers of high similarity analogs.

The number of new peptides with $CBD_{MXFP} \leq 300$ to the target keeps growing regularly as the number of PDGA runs increase, suggesting that a very large and possibly unlimited number of analogs can be produced in each case (Figure 27f). In fact, different PDGA runs rarely generate the same compounds. This is not surprising if one considers the extremely large size of the peptide chemical space, which has a size of M^N for M building blocks assembled in an *N*-mer sequence, corresponding from 10^{16} to 10^{32} for 39 building block forming sequences of 10 to 20 residues.

Table 14. Compounds used targets for PDGA

Topology	Name	Sequence [a]	Analogs [b]	Unique [c]	Target [d]
linear	indolicidin	ILPWKWPWWPWRR	2,617,229	88 %	15/50
cyclic	tyrocidine A	Cyclo[fPFfNQYVOL]	4,148,264	98 %	9/50
polycyclic	ω-conotoxin-MVIIA	C(1)KGKGAKC(2)SRLMYDC(3)C(1)TGSC(2)RSGKC(3)	8,189,307	100%	0/50
dendritic	**G3KL**	$(KL)_8(KKL)_4(KKL)_2KKL$	462,523	85 %	16/50

[a] Free carboxy termini are carboxamide $-CONH_2$. [b] Number of unique peptides generated within $CBD_{MXFP} \leq 300$ from the target compound after 50 runs of PDGA. [c] Percentage of analogs generated which occurred in only one of the 50 PDGA runs. [d] Number of runs where the target was reached.

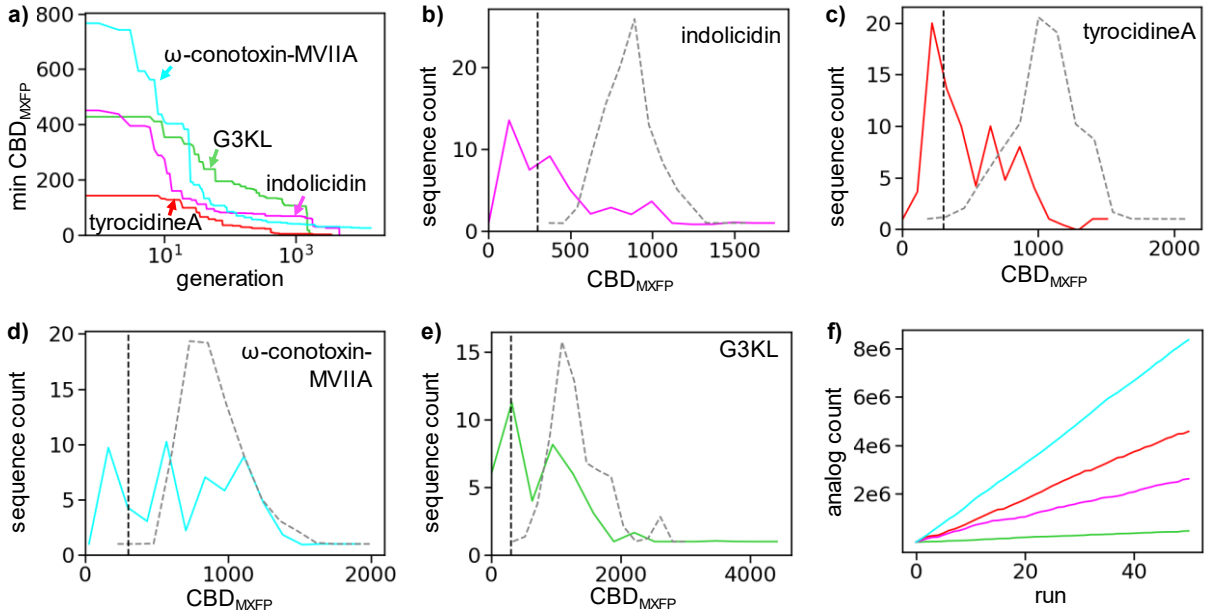


Figure 27. (**a**) Minimum CBD_{MXFP} from the target as function of generation number for an example PDGA run. (**b**) Histogram of CBD_{MXFP} distance to indolicidin for 10,000 randomly sampled linear peptides (dashed gray line) and sequences generated by PDGA (magenta). (**c**) Histogram of CBD_{MXFP} distance to tyrocidine A for 10,000 randomly sampled cyclic peptides (dashed gray line) and sequences generated by PDGA (red). (**d**) Histogram of CBD_{MXFP} distance to ω-conotoxin-MVIIA for 10,000 randomly sampled polycyclic peptides (dashed gray line) and sequences generated by PDGA (cyan). (**e**) Histogram of CBD_{MXFP} distance to peptide dendrimer G3KL for 10,000 randomly sampled peptide dendrimers (dashed gray line) and sequences generated by PDGA (green). (**f**) Cumulative plot of the unique analogs created per run.

7.2.2 Indolicidin analogs and sequence similarity comparison

In 50 runs towards indolicidin as target, each of them with a time limit of 24 h, PDGA produced 2.6 million unique peptide sequences within $CBD_{MXFP} \leq 300$ from the target, while the target sequence was found in 15 runs (Table 14). A closer analysis of the analogs showed that the structures generated by PDGA were similar in terms of size and hydrophobicity to the target indolicidin (Figure 28a, b, and c). The analogs were also similar to indolicidin when analyzing the properties of residues at specific positions without compromising the variety of the amino acid composition, reflecting the perception of pharmacophore features by MXFP (Figure 28e).

In principle, PDGA can be run using similarity measures other than CBD_{MXFP} as fitness function. In this case study, we also ran PDGA using a typical distance metric used in string

comparison, namely the Levenshtein distance (lev), to evaluate the sequence similarity to the target.[311,312] The modified algorithm (here named PDGA-lev) found its target much more efficiently than PDGA, with 49 out of 50 runs converging to the target in only 6 hours overall computing time. In this short time PDGA-lev nevertheless produced a large number (4.4 million) unique sequences with high similarity to indolicidin in terms of lev distance (lev ≤ 5), but with a larger spread in terms of MXFP similarity (Figure 28d). When compared with analogs generated using MXFP as fitness function, the PDGA-lev analogs however show a lower sequence diversity while retaining less of the indolicidin size and hydrophobicity both globally (Figure 28a, b and c) and for each amino acid position (Figure 28f).

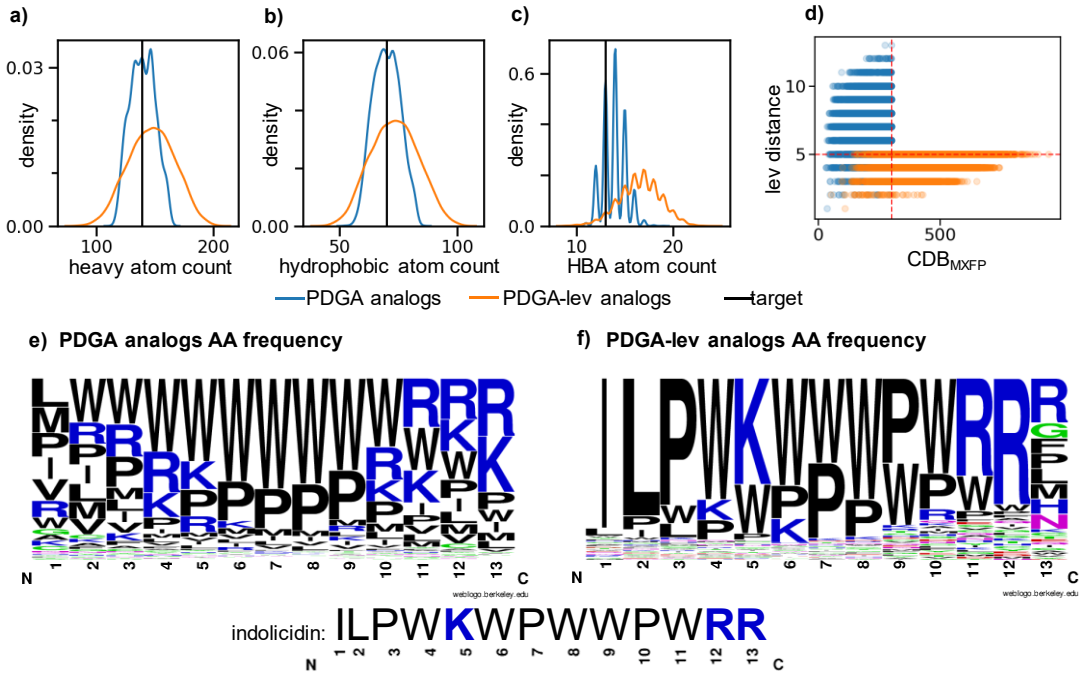


Figure 28. Physiochemical properties and amino acids composition of indolicidin analogs. Heavy atom count (**a**), hydrophobic atom count (**b**), and HBA atom count (**c**) of two 10,000 analog random subsets generated by PDGA (blue) and PDGA-lev (orange); the target values are reported as black lines. (**d**) Levenshtein distance and CBD_{MXFP} from indolicidin for 10,000 randomly picked analogs generated by PDGA (blue) and by PDGA-lev (orange); similarity threshold values are reported as red dashed lines. WebLogo[313] Amino acids (AA) frequency per position of 10,000 13-residues sequences randomly picked among the PDGA (**e**) and the PDGA-lev (**f**) analogs.

7.2.3 Cyclic and polycyclic peptides

Macrocyclic peptides are largely present in the natural realm and they are numerous among candidate and approved drugs.[314,315] Tyrocidine A (Table 14) is an antimicrobial cyclic decapeptide produced by the gram-positive *Brevibacillus brevis*, and shares structural similarities with other natural AMPs produced by the same bacteria: tyrocidine B and C, laterocidin, Gramicidin S, and loloatins A, B, C, and D.[316] When challenging our algorithm to reach tyrocidine A in 24 hours, PDGA converged to the target in 9 out of 50 runs and produced 4.1 million unique sequences with $CBD_{MXFP} \leq 300$ from tyrocidine A. In the course of these runs towards tyrocidine A, PDGA generated the well-known analog tyrocidine B. Interestingly, also the retro-sequences[317,318] of loloatin A and of tyrocidine C were retrieved as analogs by our algorithm (Table 15).

Table 15. Known analogs of tyrocidine A generated by PDGA

Analog	**Sequence** [a)]	**CBD_{MXFP}**
tyrocidine B	cyclo[fPWfNQYVOL]	67
retro-loloatin A	cyclo[VYDNfFPyLO]	157
retro-tyrocidine C	cyclo[VYQNwWPfLO]	188

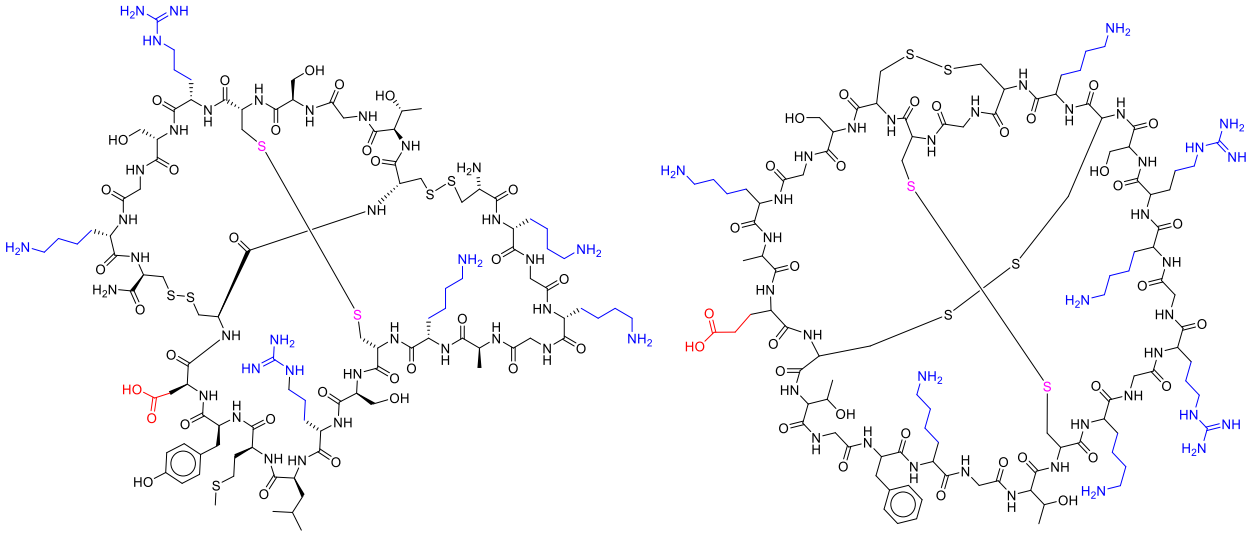
[a)] PDGA analogs do not contain any stereochemistry information, the right stereochemistry was added; O is ornithine.

For the case of ω-conotoxin-MVIIA, an analgesic 25-residue tricyclic natural peptide containing three disulfide bonds,[319] PDGA was unable to identify the target even after 72 hour runs, probably due to the polycyclic nature of the molecule. Nevertheless, the algorithm produced 8.1 million analogs with $CBD_{MXFP} \leq 300$ from the target. These analogs comprised peptides featuring a similar pattern of three cystine bridges, as well as C-to-N cyclized peptides with two or three cystine bridges and a similar overall topology. As can be appreciated by the relative positions of residues of different types in the sequences, many these analogs share the same ring topologies and distribution of cationic and anionic residues as the target conotoxin (Table 16, Figure 29).

Table 16. ω-conotoxin-MVIIA and 10 examples of its PDGA analogs.

Sequence [a)]	**CBD_{MXFP}**
ω-conotoxin-MVIIA C(1)KGKGAKC(2)SRLMYDC(3)C(1)TGSC(2)RSGKC(3)-NH_2*	-
Cyclo[C(1)KC(2)GC(3)C(2)SGKAEC(1)TGFKGTC(3) KGRGKRS]**	26
C(1)KC(2)NSTC(3)STRGKC(2)SGKLCFC(3)GRGKC(1)	28
Cyclo[C(1)KC(2)AC(3)C(2)SGKGEC(1)SGFKGTC(3)KGRGKRS]	29
Cyclo[C(1)KC(2)AC(3)C(2)SGKGEC(1)TGFKANC(3)KGRGKRS]	30
Cyclo[TSPFKGLC(1)SKSGRSC(2)EKARRC(1)GMGC(2)]	31
Cyclo[C(1)KC(2)AC(3)C(2)TGKGEC(1)SGFKGTC(3)KGRGKRC]	32
C(1)KC(2)NSAC(3)STRTKC(2)SGKASFC(3)LRGKC(1)	33
C(1)KC(2)NSTC(3)SSRGKC(2)SGKLCFC(3)VRGKC(1)	34
C(1)KC(2)NSTC(3)ATRAKC(2)SGKLSFC(3)GRGKC(1)	35
C(1)KC(2)NSTC(3)SSRAKC(2)SGKVCFC(3)LRGKC(1)	36

[a)] Disulfide bridges are indicated with matching numbers. Red = Anionic residues, blue = cationic residues. The structures of the target* and its closest analog** are shown in **Figure 29**.



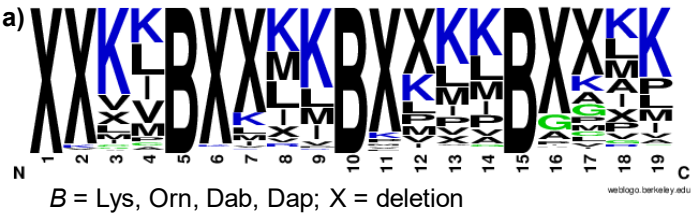
*C(1)KGKGAKC(2)SRLMYDC(3)C(1)TGSC(2)RSGKC(3)-NH2 **Cyclo[C(1)KC(3)GC(2)C(3)SGKAEC(1)TGFKGTC(2)KGRGKRS]

Figure 29. ω-conotoxin-MVIIA and the closest analog generated by PDGA.

7.2.4 Peptide dendrimers

Antimicrobial peptide dendrimer **G3KL** used as target to challenge PDGA with branched topologies belongs to a class of regularly branched peptide dendrimers, which can be prepared by solid-phase peptide synthesis and display a broad range of properties depending on the amino acid sequence and degree of branching.[270,320,321] Starting from randomly generated branched peptides with an extended time limit of 48 h, PDGA converged to **G3KL** in 16 out of 50 runs, generating 462,523 high similarity ($CBD_{MXFP} \leq 300$) analogs of the target. Among the analogs generated, 95 sequences belonged to a family of 200 high similarity analogs of **G3KL** selected from an exhaustive enumeration library consisting only of leucine and lysine among which active analogs were previously identified by synthesis and testing.[4] PDGA also generated many high similarity analogs with other residues than just leucine and lysine, thereby expanding sequence diversity while retaining high similarity to the target **G3KL** (Figure 30).

a)



B = Lys, Orn, Dab, Dap; X = deletion

b)

B	POS 6	POS 12	POS 18
Lys	62 %	38 %	30 %
Orn	30 %	37 %	15 %
Dab	5 %	18 %	10 %
Dap	3 %	7 %	45 %

c) **G3KL**:

B = Lys; X = deletion

Figure 30. (**a**) WebLogo frequency plot for 10,000 randomly selected sequences among the 319,884 PDGA analogs of **G3KL** that have three generations and a maximum 4 amino acids per generation (69 % of the **G3KL** analogs database); *B* is a branching unit and X is a deletion. (**b**) Percentage presence in the 319,884 sequences of the diamino acids branching units: lysine, ornithine, Dab, and Dap. (**c**) Reference sequence.

7.2.5 Non-peptide targets

PDGA can optimize peptides towards any molecule for which an MXFP similarity can be calculated. We therefore challenged PDGA to generate analogs of five non-peptide targets of different topologies and sizes, namely acetyl-CoA, epothilone A,[322] cholic acid, α-cyclodextrin, and a lipidated PAMAM dendrimer for siRNA transfection.[323] For each target, we ran 50 instances of PDGA for 24 hours using the same parameters as for peptides, and we set the PDGA topology to linear for acetyl CoA, to cyclic for epothilone A, cholic acid, and α-cyclodextrin, and to dendritic for the PAMAM dendrimer.

Although none of these targets could be reached and the majority of the generated sequences fell outside the similarity threshold, PDGA generated a remarkably large number of close analogs ($CBD_{MXFP} \leq 300$) in each case (Table 17, Figure 31). Note that for the smallest target molecules a relatively high number of analogs were generated more than once due to the limited possibilities found by PDGA to mimic these targets (Figure 31 h-i). Unusual amino acids such as β-alanine, γ-aminobutyric acid and hydroxyproline played a decisive role in forming high-similarity peptide analogs of the non-peptide targets (Figure 32 and Figure 52). Note that for cholic acid PDGA was run with cyclic topology, however one of the best analogs is a linear peptide which was generated because PDGA also performs linearizing mutations.

These non-peptide examples illustrate the ability of PDGA to compose a peptide structure matching the overall shape and pharmacophore of a non-peptide target molecule using amino acid building blocks only. For acetyl CoA PDGA used carboxylate side chains to mimic anionic phosphate groups, while the purine base was approximated by the indole side chain of tryptophan. In the case of epothilone A, PDGA identified macrocyclic peptides of similar size and displaying a 5-membered ring heterocycle at the correct position by selecting γ-aminobutyric acid, proline, and histidine as the most appropriate building blocks. Proline and γ-aminobutyric acid were also selected together with hydroxyproline as suitable building

blocks to mimic cholic acid with a tripeptide. α-cyclodextrin was approximated by a cyclic nonapeptide featuring hydroxyprolines, serine, threonine, glutamine and asparagine as residues with side chains featuring H-bond donors and acceptor atoms mimicking a carbohydrate.

For the case of the PAMAM dendrimer, PDGA composed a similar peptide dendrimer using amino-heptanoic acid, γ-aminobutyric acid and diaminobutyric acid branching units to mimic the extended branched structure of PAMAM. The extended hydrophobic core of the PAMAM dendrimer is approximated using aminoheptanoic acid, glycine, and a histidine featuring a 5-membered ring heterocycle matching the 1,2,3-triazole present in the target due to the click chemistry linkage used to attach the hydrophobic core to the dendrimer. Note that PDGA identified a G2 dendrimer with four end groups to mimic the PAMAM target bearing eight end-groups. Indeed, PDGA mimics the positive charges present at the branching tertiary amines and primary amine end group of PAMAM by using lysine side chains and the amino termini, which achieves a similar number of cationic groups despite of the lower generation number.

In all of these non-peptide example PDGA selects the most suitable building blocks available in the set to approximate features of the non-peptide targets. The algorithm would probably identify more similar analogs if given to choose from building blocks with better matching features such as phosphates, aromatic heterocycles, carbohydrates, and more extended lipidic chains.

Table 17. Compounds used targets for PDGA

Compound	PDGA topology [a)]	No. of analogs [b)]	Unique [c)]
acetyl-CoA	linear	13,087	24 %
epothilone A	cyclic	56,568	30 %
cholic acid	cyclic	2,694	5 %
α-cyclodextrin	cyclic	735,206	83 %
PAMAM	dendritic	114,346	97 %

[a)] Value of the input parameter topology. [b)] Number of unique peptides generated within $CBD_{MXFP} \leq 300$ from the target compound after 50 runs of PDGA. [c)] Percentage of analogs generated which occurred in only one of the 50 PDGA runs.

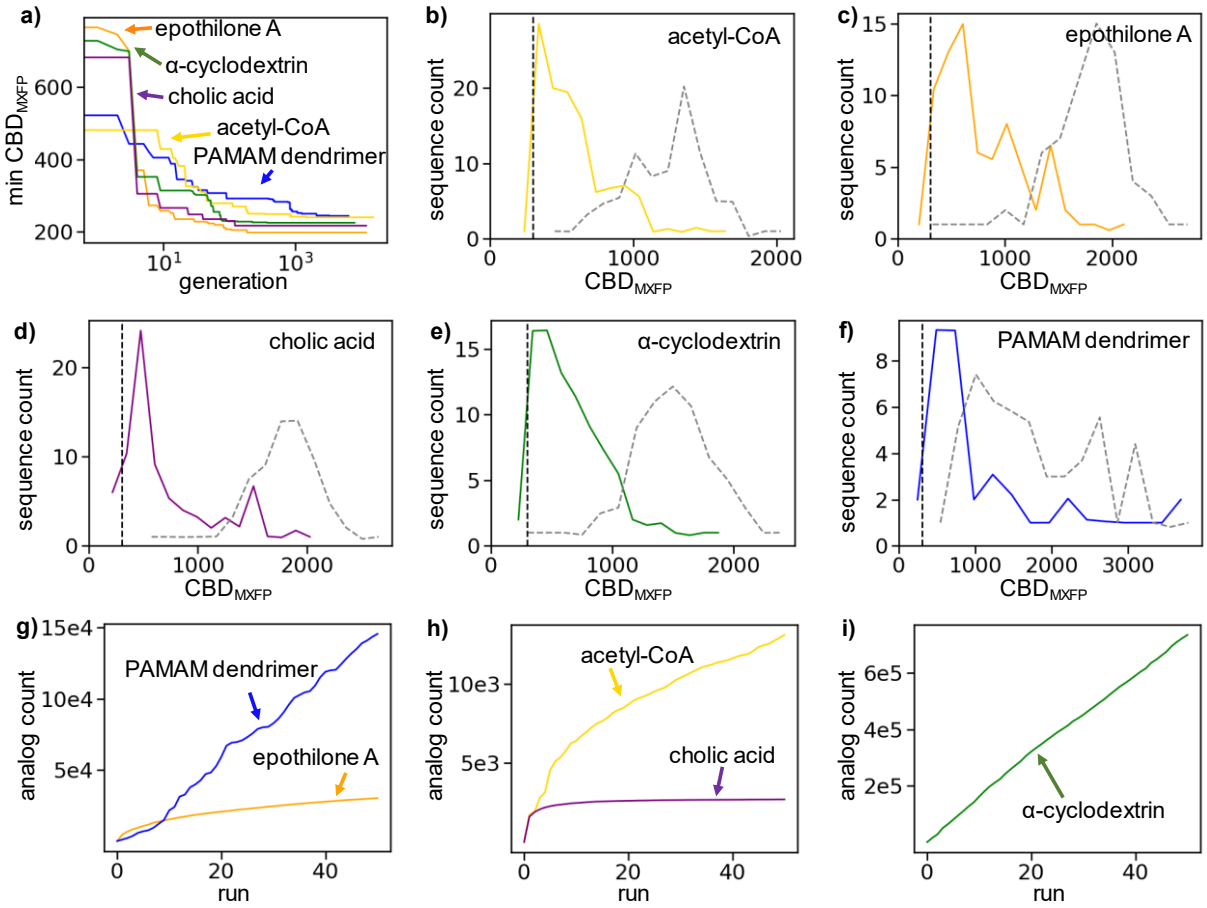


Figure 31. (**a**) Minimum CBD_{MXFP} from the target across generations for an example PDGA run. CBD_{MXFP} between the target (**b**: acetyl CoA; **c**: epothilone A; **d**: cholic acid; **e**: α-cyclodextrin; **f**: PAMAM dendrimer) and a 10,000 randomly chosen subset of the unique sequences generated by PDGA in comparison with the CBD_{MXFP} between the target and 10,000 randomly sampled sequences with the same PDGA topology (plotted as a gray dashed line). (**g-i**) Cumulative plot of the unique analogs created per run for the non-peptide targets (**g**: PAMAM dendrimer and epothilone A; **h**: acetyl-CoA and cholic acid; **i**: α-cyclodextrin).

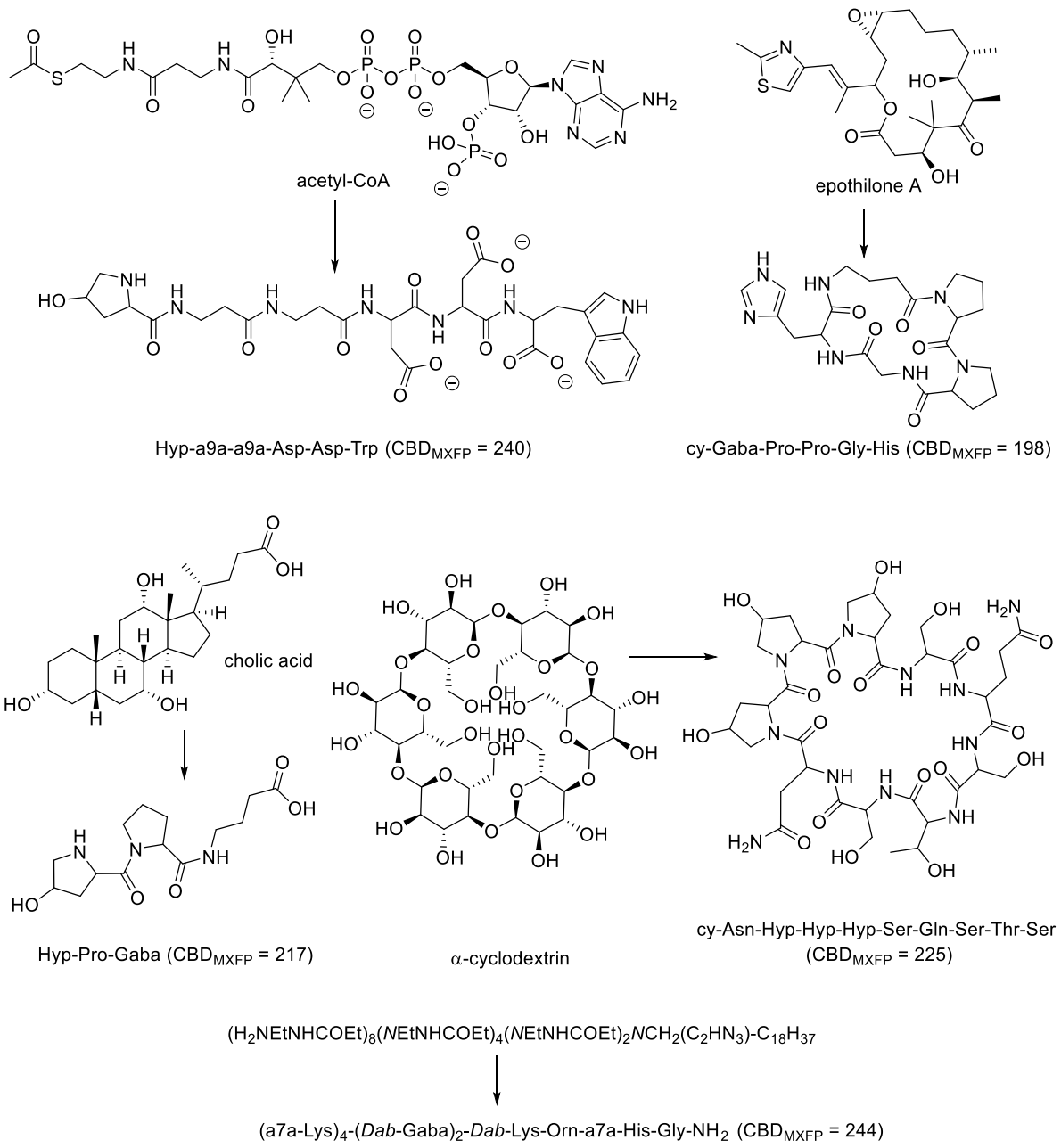


Figure 32. Structures of the non-peptidic targets acetyl CoA, cholic acid, α-cyclodextrin, and their best analogs. For dendrimer, italics indicate branching points, C_2HN_3 = 1,2,3-triazole, a7a = 7-aminoheptanoic acid, Orn = ornithine, Gaba = g-aminobutyric acid, *Dab* = branching diaminobutyric acid. See supporting information for the extended version of dendritic structures.

7.2.6 Visualizing PDGA output in chemical space

To appreciate the extent of chemical space that can be covered by PDGA, we generated a random sample of the different peptide topologies considered by the algorithm. For this study we limited our sampling to the 20 proteinogenic amino acids selected with equal probability and assembled them as linear peptides, head-to-tail cyclic and disulfide-bridged cyclic and polycyclic peptides, and regularly branched peptide dendrimers (**Table 18**). We then converted all peptide sequences to SMILES, computed 217D MXFP fingerprint, and projected the resulting dataset into 3D by principal component analysis (PCA). The resulting 3D-map, which covered 73% of data variance (PC1: 60%, PC2: 8%, PC3: 5%), was represented using Faerun, [24] a web-based application to visualize very large datasets in 3D connected to SmilesDrawer[99] to interactively draw the structural formula of the molecules corresponding to each datapoint.

The interactive Faerun map is available at http://faerun.gdb.tools including a broad range of color-coded representations. In this map molecules form a cloud resembling a helical wave with linear, cyclic and polycyclic peptides on separate sides of the first (from the left) turn of the wave and peptide dendrimers in the second turn (Figure 33a). This arrangement corresponds mostly to increasing molecule size as measured by the heavy atom count (HAC, Figure 33b). Furthermore, compounds are also distributed within the helical wave according to the fraction of aromatic atoms (Figure 33c) and to their structural intrinsic linearity (Figure 33d). Mapping the four peptides and five non-peptide target molecules and their analogs on this map illustrates how PDGA densely populates the chemical space vicinity of each target and readily stretches out of the standard peptide chemical space when challenged with non-peptide targets by using unusual amino acids (Figure 33e).

Table 18. Types of peptides considered for mapping. [a)]

topology	linear sequence length [b)]	size range (MW)	number of compounds	comment
linear [c)]	4 – 26	420.5-3328.9 Da	300,000	Linear sequences.
cyclic C-N [c) e)]	4 – 28	384.4-3424.8 Da	50,000	One cycle achieved through head-to-tail cyclization.
cyclic S-S [c) e)]	7 – 28	791.9-3371.8 Da	50,000	One cycle achieved through one disulfide.
polycyclic [c) e)]	6 – 36	629.7-3846.6 Da	250,000	Up to 4 cycles achieved through a maximum of three disulfide bonds and/or head-to-tail cyclizations.
dendritic [d)]	3 – 22	462.5-11232.6 Da	350,000	39,116 G1, 140,822 G2, and 170, 062 G3 dendrimers.

[a)] All N-termini are NH_2, all C-termini are COOH, 20 proteinogenic amino acids sampled with equal probability at variable positions. [b)] (XXXXX*B*)nXXXXX, where X is a proteogenic amino acid position with 50% chance of being empty and B is the position of a branching lysine which marks a doubling of the peptide chain. [c)] B probability of being empty = 100%, n = 5. [d)] B probability of being empty = 20%, n = 3. [e)] the sequence was cyclized head-to-tail and/or from 1 to 3 pairs of cyclized cysteines were inserted.

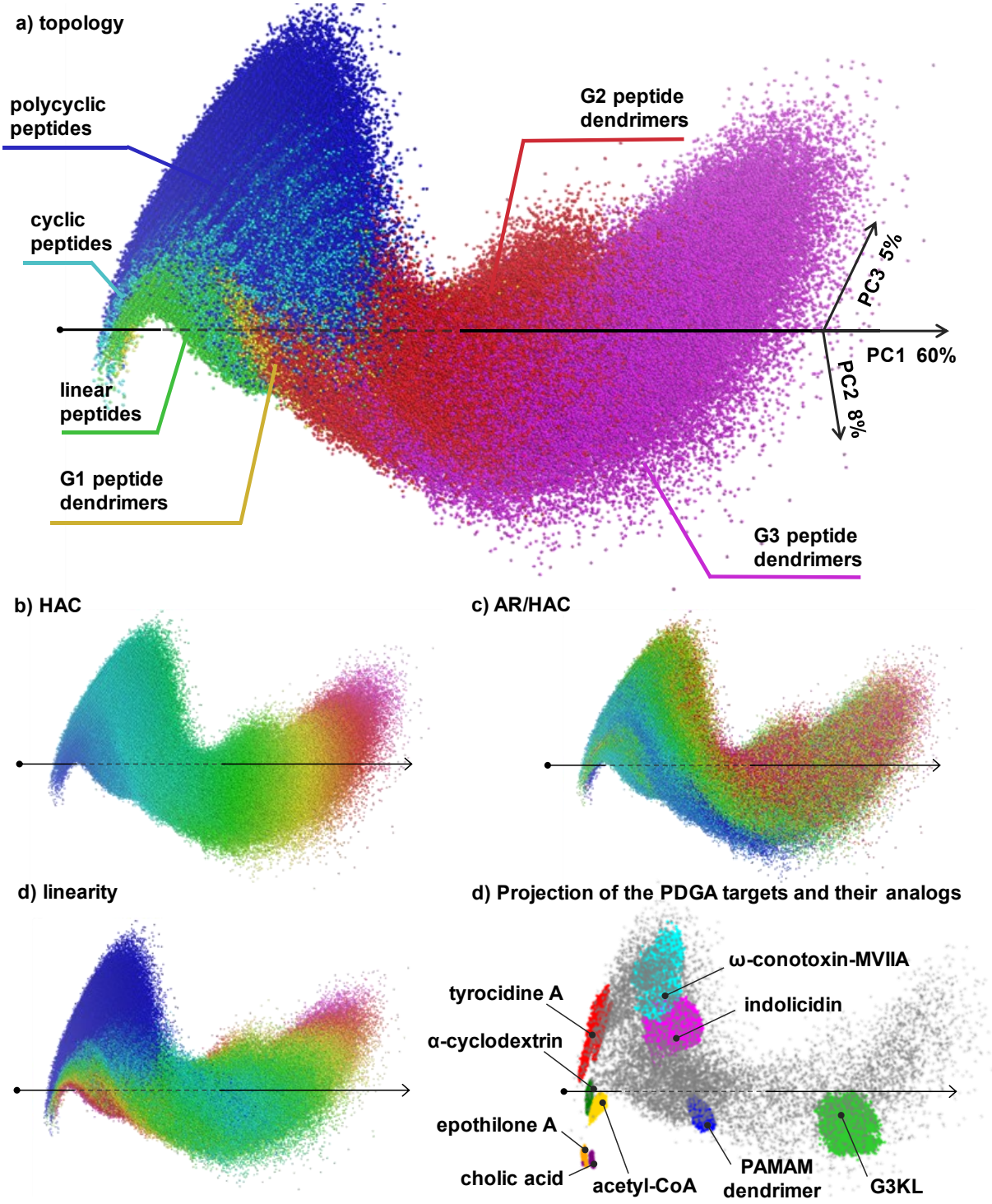


Figure 33. Peptide chemical space as visualized interactively at http://faerun.gdb.tools. PCA of MXFP values for 1 million randomly generated peptide sequences visualized in Faerun and color-coded with sequence topology (**a**, N-C and SS cyclic are grouped), HAC (**b**), AR/HAC (**c**), and linearity. (**d**) Projection of 10,000 analogs of indolicidin (magenta), tyrocidine A (red), ω-conotoxin-MVIIA (cyan), G3KL (light green), acetyl-CoA (yellow), epothilone A (orange), cholic acid (purple), α-cyclodextrin (dark green), and of the PAMAM dendrimer (blue) on a random subset of the MXFP peptide chemical space map (gray). The MXFP values of the targets are projected and reported in black.

7.2.7 Properties of PDGA analogs

The map of chemical space above illustrates the fact that PDGA analogs are very close to their targets in terms of MXFP similarity, which is somewhat trivial since MXFP was the fitness function. The abundance of high similarity analogs nevertheless illustrates the important finding that the genetic algorithm succeeds not only with linear sequences but also with cyclic, polycyclic and dendritic peptides, which is not trivial since the algorithm operates on a linear representation while the fitness function is computed on the SMILES of the actual molecules featuring the complete topology of the peptide chain. The MXFP similarity calculation is much more complex than the linear sequence comparison exemplified with indolicidin but allows more substantial sequence variations while retaining the overall molecular properties of the target and is necessary to identify peptide analogs of non-peptide targets.

The central question is whether matching the target in terms of overall molecular shape and pharmacophores by featuring a comparable size and distribution of functional groups, which are the features selected by MXFP similarity, is sufficient for a peptide to share the biological properties of the target. For indolicidin, tyrocidine A and peptide dendrimer **G3KL** which are membrane disruptive antimicrobial compounds, these features are indeed important for activity. For such membrane disruptive compounds computational designs related to our PDGA have been shown to perform well in multiple cases.[4–6,236,252,256,264,299–304] Here we found a known active analog of tyrocidine A among the high similarity analogs generated by PDGA, providing proof-of-principle of our approach (Table 15). On the other hand, the situation is expected to be more difficult when considering analogs of bioactive molecules that act by very specific interactions, such as the polycyclic conotoxins, where computational designs have not been demonstrated to date. The generated conotoxin analogs would furthermore be synthetically challenging due to the presence of multiple disulfide bridges.

For the case of the non-peptide targets, whether any of the peptide analogs generated by PDGA can display a biological activity similar to their target is probably case dependent. One would not expect any peptide mimic of acetyl CoA or epothilone A to match their activity which depend on specific reactivities and binding interactions with enzymes (AcCoA) respectively tubulin (epothilone A). On the other hand, the tripeptide analog of cholic acid identified by PDGA might display detergent properties related to bile acids. Similarly, the macrocyclic nonapeptide analog proposed as analog of α-cyclodextrin might feature comparable supramolecular complexation properties. In the case of the PAMAM dendrimer reported as an siRNA transfection reagent, we recently showed that peptide dendrimers related to those proposed by PDGA indeed display remarkable transfection reactivities.[324]

7.3 Conclusion

In contrast to previous computational approaches for peptide design which were limited to the specific case of linear peptides, we showed here that one can generate peptide analogs of any large molecule of interest, including both peptides and non-peptides, using a genetic algorithm. Our algorithm PDGA operates with peptides of diverse topologies of the peptide chain and diverse amino acid building blocks and selects molecules based on molecular shape and pharmacophore similarity to the target as fitness function. PDGA uses the Manhattan distance of the 217D atom-pair fingerprint MXFP for selection, however the algorithm can also operate using any other type of similarity measure as long as it is based on descriptors encoding information on the relative position of functional groups in a molecule, as exemplified here with the Levenshtein distance in the linear peptide example indolicidin. The pertinence of the analogs generated by PDGA is supported by the presence of known actives among these high similarity molecules in the case of macrocyclic peptide tyrocidine A and peptide dendrimer **G3KL**.

The PDGA application is versatile. On the one end through additional building blocks and cyclization types it is possible to further expand the size of the explored chemical space. On the other end selecting building blocks allows the exploration of specific chemical subspaces. The extent of peptide chemical space accessible by PDGA was illustrated by an interactive map of one million peptide structures. PDGA should be generally useful to generate peptides at any location in chemical space.

7.4 Methods

7.4.1 Converting peptide sequences to SMILES

The conversion of the peptide sequences to SMILES is performed by the Peptide Design Genetic Algorithm (PDGA). PDGA recognizes as inputs sequences of amino acids represented with their three-letter codes and separated by hyphens, with the N-terminus on the left and C-terminus on the right, considering sequences that can in principle be prepared by solid-phase peptide synthesis. For cyclic peptides, cyclization between C- and N-terminus is specified by adding “cy-” at the beginning of the sequence. Cyclization by disulfide bond is introduced by double insertion of a pair of bridged cysteines indexed with matching numbers. For peptide dendrimers, the appearance of a branching diamino acid in the sequence, marked by specific building blocks, implies that the peptide chain extending at left of the branching residue is doubled, and the total chain multiplicity depends on the number of branching residues present in the sequence. In the current implementation, the available building blocks are natural amino acids, defined with their standard three-letter code, and a selection of non-natural amino acids with assigned three-letter codes (Table 24).

The three-letter codes are translated to single characters that correspond to the one letter code symbol for the natural amino acids and are arbitrary for the other building blocks (Table

24). Starting at the sequence C-terminus (right end of the line notation), the SMILES corresponding to the first building block is transformed into an RDKit[123] molecule object and treated as structure seed. Then, each symbol is transformed into an RDKit molecule object and sequentially attached to the growing structure until the N-terminal is reached. C-terminus and N-terminus of each building block, sulfur of cyclized cysteines, and di-amines of branching units are flagged in the respective SMILES so that it could be tracked where amide and disulfide bonds must take place. Once the entire sequence is processed, the resulting RDKit molecule object is converted back into its SMILES string.

7.4.2 MXFP calculation – SMARTS

MXFP first assigns each atom to one or more of the following seven categories: non-hydrogen atom, hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, positively charged, and negatively charged. For each atom pair within each category, the topological distance, which is the shortest path between the two atoms counted in bonds, is converted to gaussian of width 18% centered on the distance itself. The sum of atom pair gaussians in each category are sampled at the following 31 distances 0, 1, 2, 3, 4, 5, 6, 7.1, 8.4, 9.9, 11.6, 13.7, 16.2, 19.1, 22.6, 26.6, 31.4, 37.1, 43.7, 51.6, 60.9, 71.8, 84.8, 100.0, 118.0, 139.3, 164.4, 193.9, 228.9, 270.0, and 318.7 bonds, which results in 217 distance values forming the fingerprints after normalization. Details of the calculation and formula have been reported previously.[14] In this work, MXFP was adapted to work faster on peptides by assigning hydrogen bond acceptor (HBA)/donor (HBD) properties and formal charges using SMARTS (Table 25). The rest of the properties were assigned as previously described, with 0.5 as scale factor for the categories HAC and Aromatic.

7.4.3 Genetic algorithm workflow

PDGA takes six input parameters: population size (N), mutation rate (α), generation gap (β), target sequence (accepted building blocks and their meaning, Table 24), similarity threshold (ST), and topology (TP: linear, cyclic, or dendritic). The algorithm produces N random sequences and assigns them a survival probability according to the reciprocal Manhattan distance between their MXFP to that of the target sequence.

Then, the new generation of peptides sequences is formed in the 4 steps reported below. 1) The best $N(1-\beta)$ sequences (survivors) are kept unchanged. 2) $N\beta$ new sequences are produced. If TP is linear or dendritic, 90% of the new sequences are created through crossover. Meaning that two sequences of the previous generation are randomly picked according to a distribution based on their survival probability and the first half of the first one is merged with the second half of the second to give a "child" sequence. The remaining 10% is randomly generated (see section 7.4.5). When TP is cyclic, only 40% of the new sequences come from crossover, 50% are picked from the previous generation sequences according to their survival probability, and the remaining 10% are generated randomly. 3) the new sequences ($Nnew$) undergo the mutation process: one of the available mutations is randomly chosen, and $\alpha Nnew$ sequences are picked and mutated for each allowed mutation. For all topologies, PDGA performs single point mutations and insertions of amino acids, and each position can undergo single point deletion. If TP is dendritic, PDGA also performs insertion and mutation of branching units; mutation can affect both the building block type and its position (±1). If TP is cyclic, PDGA performs additional "cyclization/linearization" type mutations: amide bond head-to-tail cyclization/linearization, insertion/deletion of cyclized cysteines, and transformation of amide bond head-to-tail cyclization to a couple of cyclized cysteines at the beginning and at the end of the sequence. 4) The survivors are merged with the new sequences into the new generation.

Once formed, the new generation of peptide sequences is then evaluated again, and the cycle continues until CBD_{MXFP} 0 is found for ten times or the time limit is reached. PDGA writes out all generated sequences in a "generations" file, which is updated each generation, and every sequence with CBD_{MXFP} from the target MXFP lower than the *ST* in a "results" file. The PDGA code is open source and freely accessible at https://github.com/reymond-group/PeptideDesignGA.

7.4.4 PDGA runs and reproducibility

Each PDGA use reported in this work has been repeated 50 times with different random seeds (integers from 1 to 50). In the ω-conotoxin and the **G3KL** examples, PDGA time limit was set to 72 and 48 hours, respectively; while in the other case studies PDGA was run for 24 hours. PDGA-lev was run for 6 hours, but most instances found the target in few minutes. Settings, input parameters, and the excluded building blocks are reported in Table 26. Regarding panel a of Figure 30, the reported PDGA instances have been ran with seed 7, 6, 11, 1, 1, and 37 for indolicidin, tyrocidine A, ω-conotoxin-MVIIA, **G3KL**, epothilone A, and the PAMAM dendrimer respectively.

7.4.5 Random generation of sequences

The random generation of peptide sequences is performed by the PDGA. To construct the peptide chemical space map, 350,000 peptide dendrimer sequences were generated with "XXXXX*B*XXXXX*B*XXXXX*B*XXXXX" composition, where X is a natural amino acid, *B* is a branching lysine, and the latter marks a doubling of the peptide chain at its left. All amino acids and branching residue have an equal probability of being picked in their respective positions. To allow sequences of different lengths, both X and *B* positions can be empty with a probability of 50% and 20%, respectively. If one branching unit position is empty, the neighbor generations are merged; therefore, the maximum generation length increases when

the number of generations decreases. At its maximum extent, this can lead to no branching unit and linear sequences of 20 amino acids maximum length that were filtered out.

650,000 more peptide sequences were generated with no branching units and a maximum length of 30 amino acids. Then 350,000 of these linear sequences were cyclized using amide bond head-to-tail cyclization and cysteine disulfide bonds. The sequences were divided in seven groups of 50,000 peptides and each group was cyclized with 1 S-S bond, 2 S-S bonds, 3 S-S bonds, amide bond head-to-tail, amide bond head-to-tail and 1 S-S bond, amide bond head-to-tail and 2 S-S bonds, or amide bond head-to-tail and 3 S-S bonds, respectively. To add a disulfide bond, two cyclized cysteines were randomly inserted into the sequence. The cyclization through disulfide bond is not always possible: if two cyclized cysteines are inserted next to each other, they were removed, and the cyclization did not take place; to avoid linear sequences the cysteines insertion was repeated until the desired number of cycles was present in the sequence.

According to their topology, PDGA first generation sequences and the random component of each new generation were created as described above. In this case, linearity was kept and a selection of non-natural building blocks was allowed (see Table 26 for the building blocks used in each case study and Table 24 for their meaning). To compare the PDGA results with random peptides (Figure 27 panels b-e and Figure 31 b-f), 10,000 sequences for each topology were randomly picked among the one generated for the map.

7.4.6 Property calculation

The peptide chemical space can be navigated using 16 different properties: heavy atom count (HAC), hydrophobic atom count and fraction, aromatic atom count and fraction, H-bond donor and acceptor atom count and fraction, positive atom count and fraction, negative atom count and fraction, linearity, number of branching units and sequence topology.

Linearity is a measure of the MXFP values difference between the considered sequence and the linear alkane with the same HAC. [14] The number of branching units is the count of cyclized cysteines (Cys1, Cys2, Cy3) and branching lysine (BLys) in the sequence. Sequence topology describes the possible peptide structure and can be linear, cyclic, polycyclic, or a G1, G2, or G3 dendrimer peptide. All the other properties are calculated by MXFP using ChemAxon[325] plugins.

7.4.7 PDGA analogs visualization in the peptide chemical space map

The visualization of the PDGA analogs in the peptide chemical space was done by projecting the MXFP values of the generated sequences with $CBD_{MXFP} \leq 300$ on the PCA of a 10,000 molecules random subset of the peptide chemical space. As in Faerun the PCA was performed using scikit-learn. [162]

Chapter Eight – Machine Learning Designs Non-Hemolytic Antimicrobial Peptides

This work is based on the peer-reviewed publication:

Capecchi, A.*; Cai, X.*; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. 2021. https://doi.org/10.1039/D1SC01713F.

*These authors contributed equally to the publication.

Peptide synthesis and tests were performed by Xingguang Cai and Hippolyte Personne. Molecular dynamics simulations were performed by Hippolyte Personne.

Abstract

Machine learning (ML) consists of the recognition of patterns from training data and offers the opportunity to exploit large structure-activity databases for drug design. In the area of peptide drugs, ML is mostly being tested to design antimicrobial peptides (AMPs), a class of biomolecules potentially useful to fight multidrug-resistant bacteria. ML models have successfully identified membrane disruptive amphiphilic AMPs, however mostly without addressing the associated toxicity to human red blood cells. Here we trained recurrent neural networks (RNN) with data from DBAASP (Database of Antimicrobial Activity and Structure

of Peptides) to design short non-hemolytic AMPs. Synthesis and testing of 28 generated peptides, each at least 5 mutations away from training data, allowed us to identify eight new non-hemolytic AMPs against *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, and methicillin-resistant *Staphylococcus aureus* (MRSA). These results show that machine learning (ML) can be used to design new non-hemolytic AMPs.

8.1 Introduction

Machine learning (ML) is a part of artificial intelligence consisting of using algorithms to recognize patterns in training data. In the context of computer-aided drug discovery,[31,326] ML allows one to exploit experimental structure-activity data on known drugs to generate new molecules and predict their properties and activities.[294,295,327] Generating new molecules is commonly a two-step approach that requires first a more general training and then a fine-tuning towards a specific set of characteristics. The fine-tuning of a generative ML model can be achieved with transfer learning (TL), which is essentially a second learning of a prior generative model with a smaller set of compounds.[328]

In the area of computational peptide design,[23,329] ML models for generation and activity classification can readily be trained with structure-activity data using the linear sequence of amino acids as input for the peptide structure. Efforts to develop and test ML for peptide design mostly focus on antimicrobial peptides (AMPs)[330,236] because relatively large structure-activity databases are available in the public domain.[331–335,278,336] Antimicrobial peptides (AMPs) are synthesized by microorganisms, plants, and animals as a defense against bacterial predators innate immunity. They often show good activity against multidrug-resistant bacteria, thereby offering an opportunity to address this global public health threat.[337–340]

Most AMPs are polycationic and act by disrupting bacterial membranes, usually by folding into an amphiphilic α-helix at the membrane surface,[269,341] a mechanism against which resistance is not easily obtained and which has been used broadly to guide the design of new AMPs. Unfortunately, designing amphiphilicity often results in compounds lacking selectivity against eukaryotic membranes and showing hemolytic properties, which strongly limits their use.[342] In principle, ML should be optimally suited to address this challenge by training models with data on AMPs with annotated hemolysis data.

Several ML models for AMP *de novo* design have been reported so far, and they range from classifiers for AMPs prediction applied to select sequences from randomly generated, existing, or genome derived libraries,[264,343–349,261] to standalone generative models,[267] to a combination of both generative models and classifiers.[268,350,351] Furthermore, ML has also been used in combination with evolutionary algorithms for the optimization of AMPs.[256,352] However, only two of the discussed studies considered both activity and hemolysis in the design of novel AMPs,[348,351] reflecting the challenge of avoiding hemolysis in designing AMPs and highlighting the importance of its further investigation.

Here we considered the use of ML for AMP design considering activity and hemolysis by training our models on sets of active, inactive, hemolytic, and non-hemolytic sequences derived from reported activity data. We also aimed to validate if ML can be used to identify new AMPs by testing only sequences substantially different from known AMPs. Starting with sequence information and antimicrobial and hemolysis data from DBAASP (Database of Antimicrobial Activity and Structure of Peptides),[332] which contains manually curated information on activity values and hemolysis behavior, we trained a combination of generative and predictive recurrent neural networks (RNN). To generate peptide sequences, we trained a generative model and we fine-tuned it using TL to target three problematic and often drug-resistant pathogens: the Gram-negative *Pseudomonas aeruginosa* and *Acinetobacter baumannii,* and the Gram-positive *Staphylococcus aureus*. Furthermore, to select non-hemolytic AMPs among the generated sequences, we implemented two RNN classifiers to predict antimicrobial activity and hemolysis. Our combination of supervised and unsupervised learning to design non-hemolytic AMPs is unprecedented, and it allowed us to maximize the use of highly curated data on antimicrobial activity and hemolysis. Synthesis and testing of twenty-eight of the generated and selected sequences resulted in twelve new active AMPs,

eight of which were also non-hemolytic. Detailed characterization of the best two peptides showed that they are typical α-helical membrane disruptive AMPs.

8.2 Results and Discussion

8.2.1 Machine learning

8.2.1.1 DBAASP

DBAASP contains peptides annotated with activity values, and when known, with their hemolytic behavior. This allowed us to obtain reliable AMP activity and hemolysis data. With a threshold of 32 μg/mL and 10 μM, we identified 4,774 active and 1,867 inactive linear peptides. Additionally, we considered the DBAASP peptides reported to cause less than 20% hemolysis at a concentration of at least 50 μM as non-hemolytic and the peptides reported to cause more than 20% hemolysis at any concentration as hemolytic, which resulted in 1,319 hemolytic and 943 non-hemolytic linear peptide sequences. Finally, we extracted 339 non-hemolytic peptides active against the Gram-negative bacteria *P. aeruginosa* and/or *A. baumannii* and 458 non-hemolytic peptides active against the Gram-positive bacterium *S. aureus*.

8.2.1.2 Generative models

Alone, the 339 and 458 non-hemolytic AMPs active, respectively, against *P. aeruginosa* and/or *A. baumannii* and *S. aureus* are not enough to directly train a generative model able to design a diverse set of novel AMPs. To overcome the challenge posed by the scarcity of data points on specific strains in the DBAASP, we first trained a general generative model on the entire DBAASP, and then we fine-tuned it with the smaller subset of AMPs with reported hemolysis data and specific activity (Figure 34a). The 4,774 active peptides in the DBAASP were divided

into a training and a test set, and the training set was used to train an RNN generative model to produce AMPs (prior model).

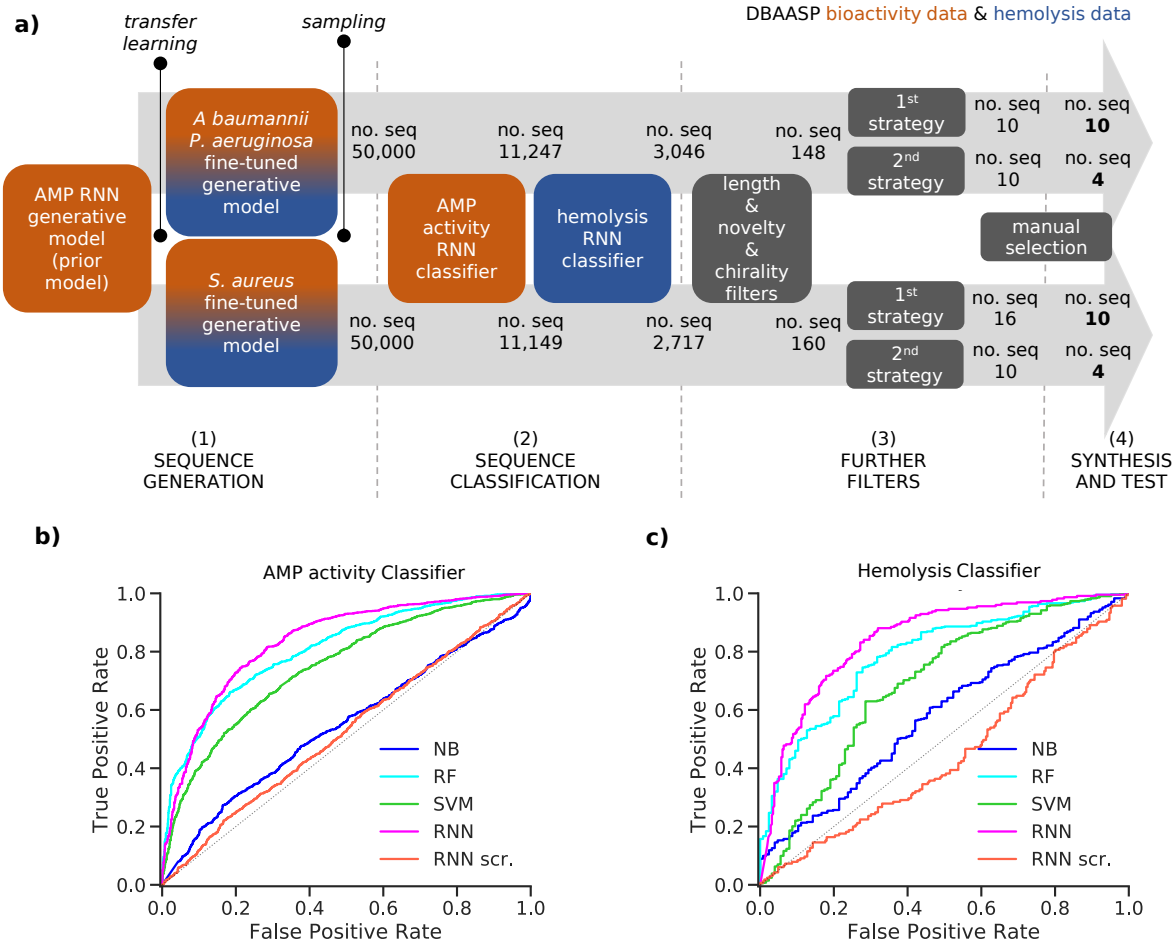


Figure 34. (a) Strategy schematic. An AMP RNN generative model, an AMP activity classifier, and a hemolysis RNN classifier were trained using activity (orange) and hemolysis (blue) data from DBAASP. (1) Two copies of the AMP RNN generative model (prior model) were transferred learned using active and non-hemolytic peptides against specific strains: P. aeruginosa/A. baumannii and S. aureus, respectively. (2) The fine-tuned models were sampled, and the generated sequences were first classified using the RNN AMP activity classifier and then the RNN hemolysis classifier. (3) The selected sequences were further filtered to obtain short peptides of maximum 15 residues with at least five mutations from the sequences in DBAASP and no D amino acids. Then two different selection strategies were used. In the first selection strategy (1st strategy) we used the calculated amphiphilicity of the sequences to further filter them, and we clustered the selected ones. In the second selection strategy (2nd strategy) we select at random 10 sequences. (4) Finally, the 28 chosen sequences were synthesized and tested. (b) ROC curves of the test set for the NB, RF, SVM, RNN, and RNN with scrambled labels (RNN scr.) models for the AMP activity (b) and hemolysis (c) classification tasks. The probabilistic prediction values were converted into binary classification values using a threshold of 0.5.

Subsequently, two generative models were derived by fine-tuning the prior model with TL using two smaller sets of sequences with a specific activity and known non-hemolytic behavior: (i) the 242 non-hemolytic peptide sequences present in the training set of the prior model and active against the Gram-negative *P. aeruginosa* and/or *A. baumannii* and (ii) the 321 non-hemolytic sequences present in the training set of the prior model and active against the Gram-positive *S. aureus*. Interestingly, 170 peptides were common to both sets (see methods section 8.4.1 and 8.4.5 for details).

To avoid overfitting, the prior and the two fine-tuned generative models were trained with the respective training sets until the probability of generating the related test sets reached their maximum value. For the fined-tuned models, the 97 and 137 sequences active, respectively, against *P. aeruginosa* /*A. baumannii* and *S. aureus*, which were present in the test set of the prior model, were used as test set. We then sampled 50,000 peptide sequences from each of the two fine-tuned models. The percentage of unique sampled sequences was 82.8% for the *P. aeruginosa* /*A. baumannii* model and 82.3% for the *S. aureus* model. Furthermore, in both cases over 99% of the sampled sequences were not present in the corresponding training set used for transfer learning due to our attention in avoiding overfitting. The high percentage of uniqueness and the novelty of the generated sequences within the 50,000 samples showed that our fine-tuned models were capable of generating new and diverse sequences. This allowed us to proceed in our analysis with a relatively small and manageable number of candidate peptide sequences.

8.2.1.3 Classifiers

To assess the capabilities of the prior model and to predict the AMP activity of the generated peptide sequences, we implemented a NB (Naive Bayes), an SVM (Support Vector Model), a RF (Random Forest), and an RNN AMP activity classifiers. The DBAASP active compounds in the same training/test split used for the prior model were used as positive class. As negative

class, we used an equally sized set of inactive sequences dived in training and test sets. The inactive sequences consisted of all inactive sequences in DBAASP and additional sequences generated by scrambling active peptides and by fragmenting SwissProt entries. As a baseline and to make sure that the performance of the RNN model was due to a trend in the data and not to an artifact, an RNN activity classifier with the same architecture but trained with scrambled labels was implemented. The models were trained using the training set, and their performances were evaluated using the test set (Figure 34b, Table 27). The RNN activity classifier performed best across all computed metrics (ROC AUC = 0.84, accuracy = 0.76, precision = 0.74, recall = 0.80, F1 score = 0.77, MCC = and 0.53) and was selected for further investigation.

To account for non-hemolytic behavior, a second classifier to distinguish between hemolytic and non-hemolytic sequences was trained. In this case, the DBAASP entries with hemolysis annotation were used to train the models. Non-hemolytic sequences were considered as the positive class and hemolytic sequences as the negative class. Being the sequences with hemolysis data a subset of the ones having activity data, we used the same training/test split used for the activity classifier (for details refer to method section 8.4.1). Similar to the AMP activity classification discussed above, an RNN classifier with scrambled labels (baseline), NB, SVM, RF, and RNN classifiers were trained with the training set and evaluated for the hemolysis task with the test set. As for the activity classifier discussed above, the RNN classifier had the best overall performance for hemolysis prediction (ROC AUC = 0.87, accuracy = 0.76, precision = 0.70, recall = 0.76, F1 score = 0.73, MCC = 0.52) and was selected for further study (Figure 34c, Table 27).

To increase the precision of the RNN AMP activity and RNN hemolysis classifiers, we raised the threshold used to transform their probabilistic output to a binary classification from 0.5 to over 0.95 for both classifiers (refer to methods 8.4.6 for details). This resulted in an

adjusted precision of 0.91 and 0.84 for the RNN AMP activity classifier and the RNN hemolysis classifier, respectively. Therefore, when considering the antimicrobial activity and the hemolysis behavior of a peptide sequence as two independent characteristics, we obtained a combined precision of 0.76, which means that 76% of predicted positives are expected to have antimicrobial activity and non-hemolytic properties. However, because hemolysis is a known drawback of antimicrobial peptides, non-hemolytic behavior and antimicrobial activity are likely to be inversely proportional. This is also evident when looking at the 1,786 active peptides reported in the DBAASP with a hemolysis annotation, as only 721 are reported as non-hemolytic. For this reason, a lower overall performance of the two classifiers was expected.

8.2.1.4 Sequences selection

The RNN AMP activity and hemolysis classifiers were used to filter the 50,000 sequences sampled from each of the two fine-tuned generative models, resulting in 3,046 sequences from the model fine-tuned for *P. aeruginosa* and *A. baumannii* and 2,717 from the model fine-tuned for *S. aureus* (Figure 34a). To facilitate the synthesis process, sequences longer than 15 amino acids were excluded (Figure 53a). The sequences were further filtered to ensure novelty, considering a minimum of four mutations from the test set peptides and, to further challenge our model, of five mutations from the training set peptides (Figure 53b to e). This selection criterion has not been used in previous AMP discovery approaches using ML, however, we believe it to be fundamental to avoid trivial analogs of known peptides and analogs which have already been studied within SAR analysis. Finally, since the percentage of D amino acids in the training sets of the generative model and of the classifiers was low, we decided to exclude the sequences containing D amino acids since the data would be insufficient for the model to learn features for such peptides (Figure 53f). The selection yielded 148 and 160 peptides from the *P. aeruginosa*/*A. baumannii* model and *S. aureus* model, respectively.

Then, two different strategies to further select the sequences were followed. In the first case, we used the calculated hydrophobic moment[353] and the predicted α-helix fraction as estimations of amphiphilic helix to further filter the sequences (Figure 53g) and performed clustering to diversify our selection (first selection strategy). In the second case, we randomly sampled 10 sequences out of each pool of peptides to follow the model sampling distribution (second selection strategy, see methods section 8.4.7 for details). This selection resulted in 20 peptide sequences from the *P. aeruginosa*/*A. baumannii* model and 26 peptide sequences from the *S. aureus* model. From each set, 14 peptides were chosen manually for experimental evaluation. Thanks to the applied filters and selection processes, all selected sequences were distinct from the training and test sets of both AMP activity and hemolysis classifiers in at least five positions, and to the best of our knowledge, they were not present in any peptide databases. The sequences coming from the *P. aeruginosa*/*A. baumannii* model were labeled as Gram-negative targeting compounds (GN), and the sequences selected from the *S. aureus* model were labeled as Gram-positive targeting compounds (GP).

8.2.2 Synthesis and testing

8.2.2.1 Antibacterial activity and hemolysis

We synthesized the selected 14 GN and 14 GP peptides by solid phase peptide synthesis and evaluated the activity of their HPLC-purified trifluoroacetate salts by determining minimum inhibitory concentrations (MIC) against bacteria by broth microdilution assay in Muller-Hinton medium and minimum hemolysis concentrations (MHC) on human red blood cells by serial dilution in phosphate buffer saline (**Table 19.** Synthesis and activity of generated peptides.Table 19).

Considering an activity threshold of MIC ≤ 16 µg/mL for activity and MHC ≥ 500 µg/mL for hemolysis, 9 of 14 GN peptides (64 %) turned out as actives, but only 6 of 14 GN

(43 %) were both active against *P. aeruginosa* or *A. baumannii* and non-hemolytic. By the same measure, only 3 of 14 GP peptides (21 %) were active against MRSA, and only 2 of 14 GP peptides (14 %) were also non-hemolytic. Furthermore, three of the active GN peptides were also active against MRSA, while all three active GP peptides and one GP inactive peptide were also active against *P. aeruginosa* or *A. baumannii*, and 11 out of 14 GN and 6 out of 14 GP peptides showed activity against *Escherichia coli* tested as an additional Gram-negative bacterium. Therefore, in terms of overall activity, 18 out of the 28 synthesized peptides (64 %) were active below the threshold, and 14 out of 28 (50 %) were active and non-hemolytic, which is not very much below the precision of 76 % for the combined activity/hemolysis classifier (see above).

The lack of selectivity of the generated AMPs for the bacteria they were trained on, either Gram-negative (*P. aeruginosa* and *A. baumannii*) or Gram-positive (*S. aureus*) bacteria suggested to test our AMPs in a broader context. We therefore tested the best GN (**GN1**) and the best GP (**GP1**) AMP against additional pathogenic bacteria available in our laboratory (Table 20). Both peptides were also active against ZEM-1A, which is a multidrug-resistant clinical strain of *P. aeruginosa*, but not against the related ZEM9A which is more resistant to polymyxin B, a pattern which we have observed previously with other AMPs.[5,122] **GN2** also showed good activity against *P. aeruginosa* PA14 and several mutant strains generated to be resistant to polymyxin and antimicrobial dendrimers,[272] and against *S. maltophilia*, *E. cloacae*, both Gram-negative, and to a lesser extent against *S. epidermidis* (Gram-positive), but was inactive against two different strains of *Klebsiella pneumoniae* (Gram-negative). **GP1** also showed significant activity against several of these strains, and even against the two *K. pneumoniae* strains. This extended profiling confirmed the robust activity of both AMPs but also underscored the fact that our generative models did not produce AMPs with selectivity

between Gram-negative and Gram-positive strains, reflecting the fact that many AMPs appeared as actives in both TL training sets.

Table 19. Synthesis and activity of generated peptides.

cpd [a]	Sequence [b]	*P. aeruginosa* [c] (µg/mL)	*A. baumannii* [c] (µg/mL)	MRSA [c] (µg/mL)	MHC [d] (µg/mL)	*E. coli* [c] (µg/mL)
Gram-neg. active, non-hemolytic:						
GN1	**AKRIRKLIKKIFKKI**	**4**	**4**	**16**	**>2000**	**8**
GN2	RRWKWRRKIKKWL	8	8	4	1000	16
GN3	IDKWKAAFKKIKNLF	8-16	2	8	500	16-8
GN4	LNALKKVFQKIRQGL	32	16	>64	>2000	4
GN5	KFFRKLKKLVKK	16	>64	64	>2000	64
GN6	RLRKKWRKLKKLL	32	32-16	64	2000	32-16
Gram-neg. active, hemolytic:						
GN7	KRIRKWVRRILKKL	4	4	4	250	16
GN8	LRKFWKKIRKFLKKI	8	4	4	62.5	16
GN9	KRLWKRIYRLLKK	8	8	8	250	8-4
Gram-neg. inactive:						
GN10	IRRIRKKIKKIFKKI	32	32	64	>2000	16
GN11	LRKARRLLKKLRARL	>64	32	32	>2000	32
GN12	GNWRKIVHKIKKAG	32	>64	>64	>2000	16
GN13	AGRLQKVFKVIAK	64	>64	>64	>2000	32
GN14	IHKLAKLAKNVL	>64	>64	>64	>2000	32
Gram-pos. active, non-hemolytic :						
GP1	**FLKAVKKLIPSLF**	**16**	**8-16**	**16**	**2000**	**8**
GP2	RWRWPILGRILR	8	16	16	500	16
Gram-pos. active, hemolytic:						
GP3	FLHSIGKAIGRLLR	16	16	8	250	8
Gram-pos. inactive:						
GP4	GIGAVLNVAKKLL	64	32	32	>2000	16
GP5	KVARFLKKFFR	64	64-32	32	>2000	4
GP6	LKKLWKRIIKVGR	32	32-16	64	>2000	8
GP7	ARKWRKFLKKI	>64	64	64	>2000	64-32
GP8	GRIKRIRKIIHKY	8	32	>64	>2000	32
GP9	ARKKWRKRLKKLKI	64-32	>64	>64	>2000	64-32
GP10	AKKVVKKIYKRFQK	>64	64	>64	>2000	64
GP11	ARKFRRLVKKLR	>64	>64	>64	>2000	64
GP12	LRKARRLVKKLA	>64	>64	>64	>2000	>64
GP13	KRLWKIRQRIAK	>64	>64	>64	>2000	32
GP14	LNALKKVFQKIH	>64	>64	>64	>2000	>64

a) Compounds labeled as GN were obtained from the *P. aeruginosa*/*A. baumannii* model, compounds labeled as GP were obtained from the *S. aureus* model; in both sets, compounds were ordered according to their activity and hemolysis profile; **GN2, 6, 9, 10** and **GP2, 6, 9, 11** were obtained using the second selection strategy. b) One-letter code for amino acids. All peptides are carboxamides (-$CONH_2$) at the C terminus. c) MIC was determined after incubation for 16-20 h at 37°C. d) MHC was measured on human red blood cells in 10 mM phosphate buffer saline, pH 7.4, 25°C. 0.1% Triton X-100 was used as a positive control. Highlight in green denotes MIC < 32 µg/mL towards the bacterial strains used for the design (*P. aeruginosa*/*A. baumannii* for GN and *S. aureus* for GP) or MHC ≥ 500 µg/mL.

Table 20. MIC [a)] of **GN1** and **GP1** towards further MDR and non-MDR bacterial strains.

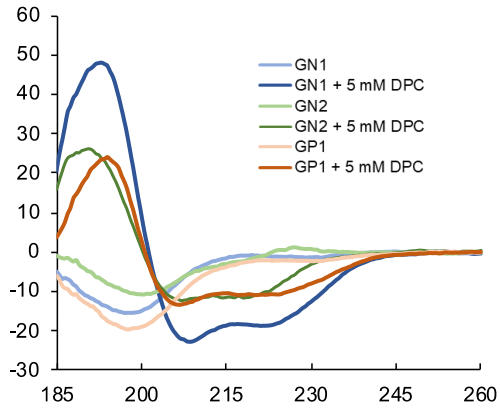
	GN1	**GP1**	Polymyxin B
P. aeruginosa ZEM-1A [b, c,)]	4	4	0.5
P. aeruginosa ZEM9A [b, c,)]	64	64	4
P. aeruginosa PA14 [c)]	2	8-16	<0.5
P. aeruginosa PA14 4.13 (*phoQ*) [c, d)]	2	8-16	1
P. aeruginosa PA14 4.18 (*pmrB*)[c, d)]	4	32-64	2
P. aeruginosa PA14 2P4 (*pmrB*)[c, d)]	8	64	2
S. maltophilia [b, c)]	4	16	0.5
E. cloacae [b, c)]	8	16-32	1
K. pneumoniae (OXA-48) [b, c)]	>64	16-32	1
K. pneumoniae NCTC148 [b, c)]	>64	32	1
B. cenocepacia [b, c)]	>64	>64	>64
S. epidermidis [b, e)]	16	16	32-64

a) The MIC was determined in Müller-Hinton medium after 16-20 h of incubation at 37 °C. Each result represents two independent experiments performed in duplicate. b) MDR strains. c) Gram-negative strains. d) Strains carrying spontaneous mutations in the indicated genes, all leading to polymyxin B resistance. e) Gram-positive strain.

8.2.2.2 α-helical folding and membrane disruption

The amino acid sequences of peptides **GN1** (15 residues, 8 cationic, 7 hydrophobic) and **GP1** (13 residues, 3 cationic, 9 hydrophobic) both had an amphiphilic composition. Circular dichroism (CD) spectra showed that both peptides were unordered in pure water but adopted an α-helical conformation in the presence of n-dodecyl phosphocholine (DPC) micelles mimicking the membrane environment. The effect was very strong with **GN1** (89 % α-helix with 5 mM DPC) and still quite strong with **GP1** (56 % α-helix with 5 mM DPC) despite the presence of a helix-breaking proline residue in its sequence and in line with the fact that this sequence passed the α-helical filter used for sequence selection. By comparison, the second most active, non-hemolytic AMP **GN1** (13 residues, 8 cationic, 5 hydrophobic) which had been selected from the RNN generator and classifiers without the α-helix filter, only showed 36 % α-helix with 5 mM DPC. Nevertheless, all three AMPs were predicted to adopt an amphiphilic arrangement of their cationic and hydrophobic side chains upon α-helical folding (Figure 35c).

a)



b)

	Helix	Strand	Turns	Unordered
GN1	16.2%	18.4%	28.7%	36.7%
GN1+DPC	89.1%	0.0%	8.1%	2.9%
GN2	9.1%	29.5%	27.4%	34.0%
GN2+DPC	36.2%	18.7%	20.5%	24.5%
GP1	17.1%	18.0%	28.9%	36.1%
GP1+DPC	55.5%	6.1%	22.3%	16.2%

c)

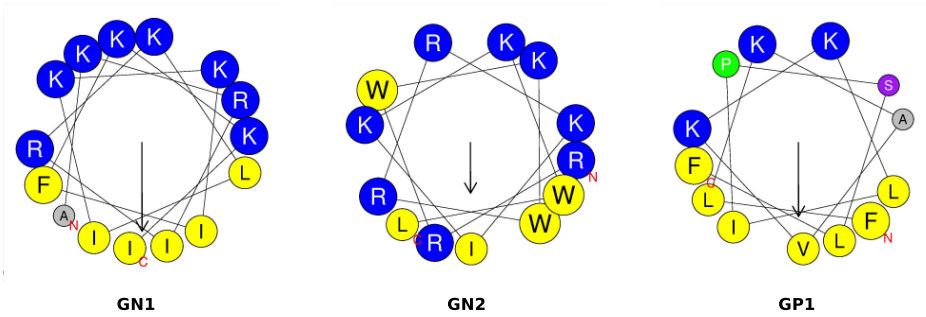


Figure 35. (**a**) CD spectra of **GN1**, **GN2**, and **GP1** recorded at 0.100 mg/mL in 10 mM phosphate buffer pH 7.4 with or without 5 mM DPC. (**b**) Extraction of percentages of secondary structure from primary CD data using DichroWeb. The Contin-LL method and reference set 4 were used. (**c**) Helix properties predicted by HeliQuest. Circle size proportional to side-chain size, blue indicates cationic residues, yellow indicates hydrophobic residues, grey indicates alanine, green indicates proline, purple indicates serine. The arrows inside each helix wheel indicates the magnitude and direction of the hydrophobic moment.

To confirm the secondary structure determined by CD, we performed MD (Molecular Dynamics) simulations for our most active peptides **GN1**, **GP1,** and **GN2** using GROMACS.[354] In each case, 250 ns simulations were performed both in water and in presence of DPC micelle. As expected, simulation in water led to the unfolding of **GN1** (Figure 36a). Interestingly, **GN1** kept a complete amphiphilic α-helix after 250 ns in presence of DPC micelle (Figure 36b, c and d), which is consistent with the 89% α-helix obtained with 5 mM DPC during the CD measurements. Similarly, **GP1** and **GN2** unfolded in water and partially folded in presence of DPC micelle (Figure 56, Figure 57). Partial α-helical conformation was observed in the case of **GP1** while interacting with the micelle, confirming the CD data and the conservation of the secondary structure despite the presence of a proline residue. Surprisingly, **GN2** unfolded and

refolded into a stable partial π-helix in contact with DPC, suggesting a stable transition state between α-helix and random coil. As both types of helices cannot be distinguished using CD, this is coherent with the helicity signal observed with 5 mM DPC. Overall, MD simulations confirmed a helical secondary structure behavior in a membrane-like environment.

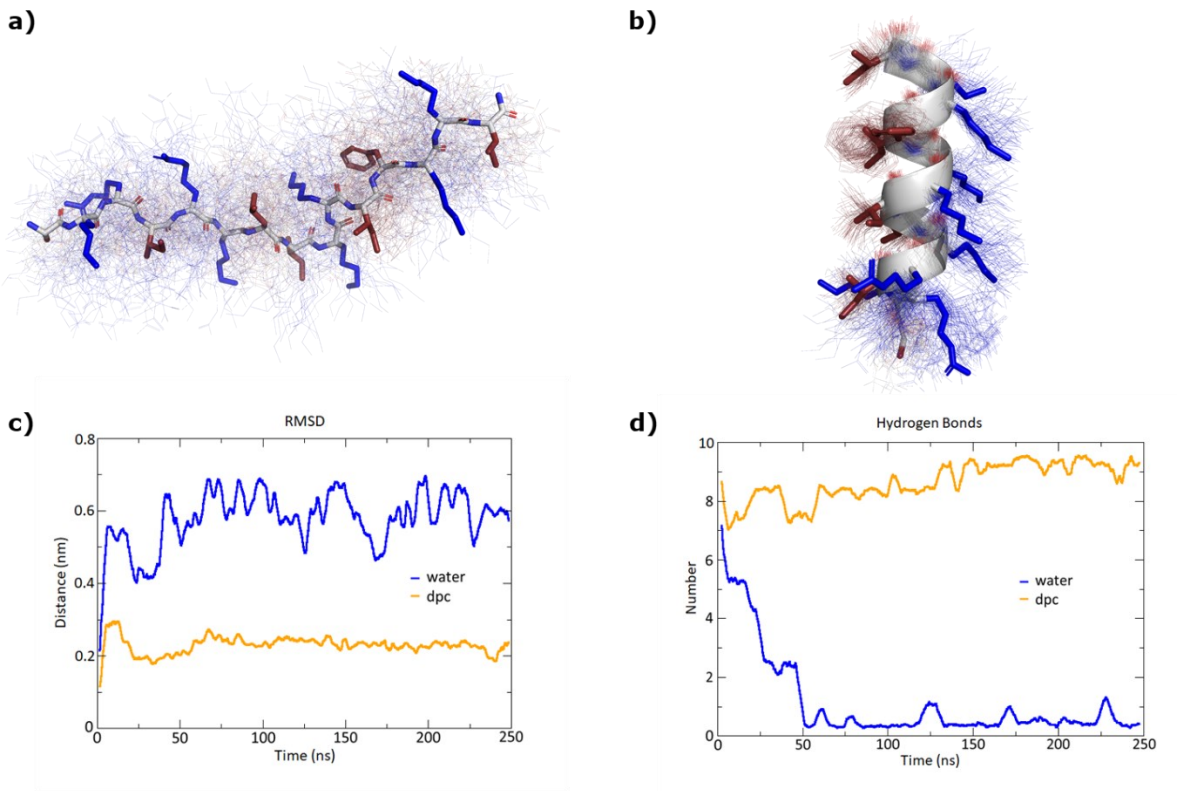


Figure 36. MD simulations of **GN1** in water and in presence of a DPC micelle over 250 ns using GROMACS. (**a**) Average structure (stick model) in water over 100 structures sampled over the last 100 ns (thin lines). Hydrophobic side chains are colored in red and cationic side chains in blue. (**b**) Average structure (cartoon model for backbone and stick model for side chains) with DPC micelle over 100 structures sampled over the last 100 ns (thin lines). (**c**) RMSD (root mean square deviation) of the peptide backbone atoms relative to the starting α-helical conformation. (**d**) Number of intramolecular hydrogen bonds. The DPC micelle was omitted for clarity.

The CD, MD, and sequence analysis above clearly pointed to membrane disruption as the probable mechanism of action for our AMPs. This hypothesis was further supported by transmission electron microscopy (TEM) imaging of bacterial cells exposed to the AMP in the case of **GN1**, which showed bacterial membrane ruptures for *P. aeruginosa*, while in the case

of *A. baumannii* the cell shape was preserved but cell contents were altered, an effect also observed with other membrane disruptive AMPs on this bacterium (Figure 37).

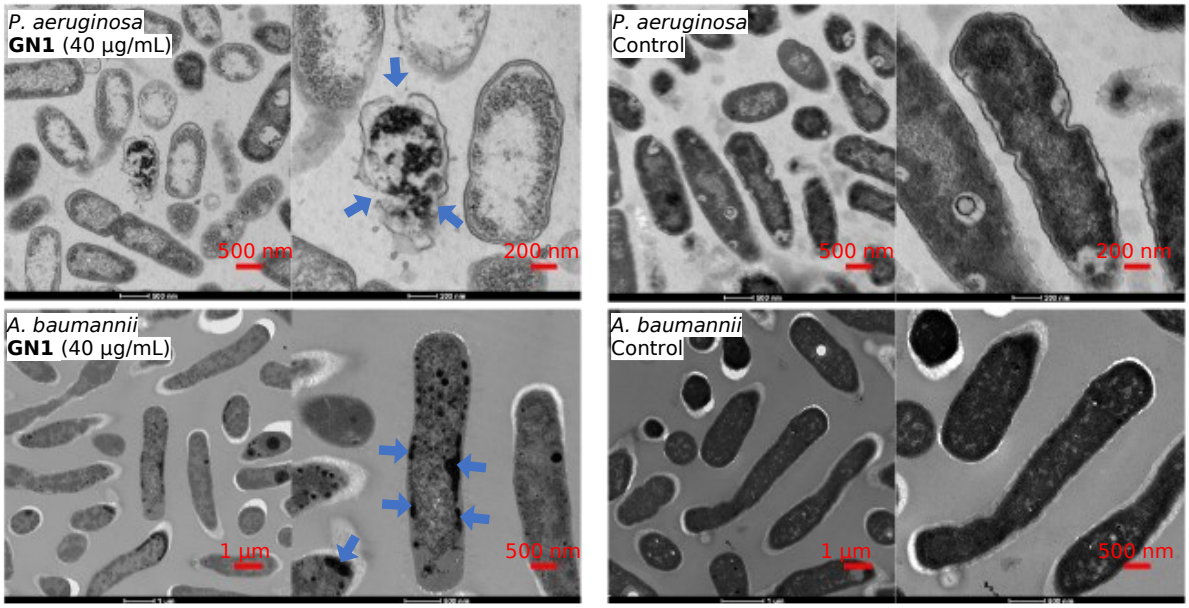


Figure 37. TEM images of *P. aeruginosa* and *A. baumannii,* after 2 hours treatment of **GN1** in MH medium. Blue arrows indicate effects on the bacteria.

8.3 Conclusion

In this work, we have demonstrated ML capable of designing non-hemolytic AMPs. We extracted a highly reliable dataset of AMPs and non-AMPs, as well as hemolytic and non-hemolytic peptides from the DBAASP, a manually curated antimicrobial peptide database. We used the data to train a generative peptide model (prior model), an AMP activity classifier, and a hemolysis classifier. Two copies of the prior model were fine-tuned using active and non-hemolytic peptides against specific strains: *P. aeruginosa*/*A. baumannii* and *S. aureus*, respectively. The fine-tuned models were sampled, and the generated sequences were filtered using the implemented classifiers, basic physicochemical properties, and novelty criteria to

obtain short peptides of maximum 15 residues with at least five mutations from the sequences in DBAASP.

Out of the 28 synthesized peptides, 12 were measured active towards the pathogens used in the design (*P. aeruginosa*/*A. baumannii* or *S. aureus*) with a MIC < 32μg/ml, which was the activity threshold selected to train our ML models, and eight of them showed low hemolysis against human blood cells with an MHC ≥ 500 μg/ml. Additionally, our best compounds **GN1** and **GP1** displayed remarkable activity also against a broader panel of pathogenic bacteria including MDR strains.

In the context of the AMPs previously discovered through a ML-guided approach,[264,268,345,348–350] **GN1** and **GP1** have a broader and overall higher activity combined with better hemolytic behavior. Two notable exceptions are the AMPs reported by Nagarajan *et. al.*[350] which have activity and hemolysis comparable to our results, and the two AMPs reported by Cherkasov *et. al.*[264] which show higher activity but a worse hemolytic behavior than our compounds. However, in both cases, hemolysis was not a design feature and the low hemolysis of the reported compounds was serendipitous. Our results indicate that ML can acquire sufficient information from known AMPs to guide the discovery of new AMPs substantially different from the training set and that ML can overcome the challenging task of designing both antimicrobial activity and non-hemolytic behavior. It should be noted that the ML approach exploiting experimental data helped us discover non-hemolytic AMPs even in the absence of a simple design rule for this property, highlighting the usefulness of ML in peptide design.

8.4 Methods

8.4.1 Datasets Preparation

All peptide sequences without intrachain bonds were downloaded from the DBAASP peptide database website (https://dbaasp.org/), resulting in a dataset of 11,805 linear peptides. Only the 9,946 sequences with free or amidated C-terminus, free or acetylated N-terminus, and containing only natural amino acids and their D-enantiomers were kept.

The targets and the activity measurements of the 9,946 sequences were extracted using the DBAASP Python API. Sequences with a registered activity measure below 10 μM, or 10,000 nM, or 32 μg/ml towards at least one reported target were labeled as active; the sequences active against *P. aeruginosa*, *A. baumannii*, or *S. aureus* were flagged. Sequences with registered activity measures above 10 μM, or 10,000 nM, or 32 μg/ml towards all reported targets were labeled as inactive; when *P. aeruginosa*, *A. baumannii,* or *S. aureus* was one of the reported targets the sequences were flagged. When present, activity against human erythrocytes was used to label the sequences as hemolytic or non-hemolytic. The concentration was normalized to μM and sequences causing less than 20% of hemolysis with a concentration equal or above 50 μM were flagged as non-hemolytic. Sequences causing more than 20% of hemolysis were flagged as hemolytic regardless of the concentration. The remaining sequences, together with the ones not having reported data against human erythrocytes, were labeled as of unknown hemolytic properties. The procedure resulted in 4,774 peptides labeled as active, 1,867 labeled as inactive, 1,319 labeled as hemolytic, and 943 labeled as non-hemolytic.

To achieve a balanced dataset for the activity classifiers, 2,907 additional inactive sequences were generated. (1) 1,453 unique sequences with the same length distribution of a randomly selected subset of the active sequences were obtained fragmenting an equally sized set of sequences randomly selected from Swissprot. (2) 1,454 unique sequences were obtained

scrambling a randomly selected subset of the active sequences. The 9,548 obtained active and inactive unique peptide sequences were divided in training and test with a 75-25 random split. In the evaluation process, the active sequences were considered as the positive class and the inactive sequences as the negative class. For the hemolysis classifier, we used the same training test split but selecting only the sequences with hemolysis data. In the evaluation, we considered the non-hemolytic sequences as the positive class and the hemolytic sequences as the negative class.

8.4.2 NB, SVM, and RF Classifiers

The NB, non-linear SVM, and RF classifiers were implemented using scikit-learn. [162] The sequences were padded to the maximum sequence length (190 residues) and tokenized as singular amino acids (or empty position), then each token was mapped to a unique number. The SVM and the RF models were optimized with a grid search to increase the ROC AUC of the test set (Table 19).

8.4.3 RNN Classifiers

The AMP activity RNN classifier and the hemolysis RNN were implemented in PyTorch.[355] The input of the implemented RNN classifiers are the tokenized and "one-hot" encoded sequences. The sequences were tokenized as singular amino acids and a start and an end tokens were added; then each token was mapped to a unique number. The resulting vector was transformed into a matrix where the number of columns is the length of the vocabulary and the number of rows was the length of the vector itself. The presence of a specific residue at each position was represented with a 1 while the rest of the matrix is filled with zeros.

The models were composed of an embedding layer, gated recurrent unit (GRU)[356] cells, and a linear transformation layer followed by a softmax function.[357] The output of the model was considered only when the last token was reached (Figure 54). The hyperparameters of the RNN

classifiers were optimized to maximize the ROC AUC of the test set (Table 27). A threshold was picked to keep the prediction of false positives below 6%. The parameters were learned using a negative log-likelihood loss[357] and a stochastic gradient descent[358] with a momentum of 0.9 and a learning rate of 0.01.

To create a baseline prediction for both RNN classifiers, a second RNN AMP activity and hemolysis classifiers (RNN AMP activity classifier scrambled labels and RNN hemolysis classifier scrambled labels) were implemented (Table 28) and trained using a different dataset, where the sequences were the same, but the activity and the hemolytic labels were randomly scrambled.

8.4.4 RNN Generative Models

A generative model was implemented in PyTorch with the same architecture of the previously described RNN activity classifier, with the exception of the dimensionality of the last linear layer which is the same size of the vocabulary (41 tokens, 41 dimensions, Figure 55). Furthermore, in this case, the output of the model was considered at every token, allowing the sequence generation. The input sequences were processed as for the RNN classifiers. The parameters of the RNN generative model were learned using negative log-likelihood loss (NLLL) and Stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.001. During the training of the generator, only the active sequences of the training set were used, but the NLLL on the test set was also monitored. The training was stopped when the NLLL of the test reached its minimum.

8.4.5 Transfer Learning

The 242 active sequences of the training set flagged against *P. aeruginosa* or *A. baumannii* and annotated as non-hemolytic were used to train again the generative model and fine-tune it against gram-negative bacteria. The 312 active sequences of the training set flagged against *S.*

aureus and annotated as non-hemolytic were used to train again the generative model and fine-tune it against gram-positive bacteria. The parameters were learned using negative log-likelihood loss (NLLL) and Stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.00001. As for the training of the prior model, the NLLL on the flagged subset of the test set, consisting of 97 for the *P. aeruginosa* or *A. baumannii* and of 137 sequences for the *S. aureus* model, was monitored and when it reached its minimum the training was stopped.

8.4.6 Sampling and Properties Calculation

50,000 sequences were sampled from each of the two transfer learned models. The Levenshtein distance (LD) from the nearest neighbor (NN) in the training and the test of both RNN classifiers was calculated using the Levenshtein Python package.[311,359] The helicity prediction was performed using SPIDER3,[360] and the helicity fraction was calculated as the number of residues predicted helical in a peptide sequence divided by the length of the sequence itself. The hydrophobic moment was calculated as described by Eisenberg et al.[353] Hemolysis and activity were predicted by the respective classifiers converting the probabilistic prediction values into binary classification using the threshold that kept the prediction of false positive below 6% (0.99205756 for the activity classifier and 0.99981695 for the hemolysis classifier).

8.4.7 Sequences Selection

The generated sequences were filtered based on multiple criteria. First, to ensure novelty, we have chosen sequences with LD > 5 from the hemolysis classifier training set sequences and LD > 4 from the hemolysis classifier test set sequences. Second, we remove all sequences that were outside the applicability domain of the hemolysis classifier. To do so, we calculated the minimum LD of every test set compound to the training set. Giving this minimum LD values we defined to applicability domain of the classifier to be the 90% quantile. This led to the exclusion of all generated sequences with a LD distance of 8 or more to the training set of the

hemolysis classifier. Only sequences up to 15 residues were selected to facilitate the synthesis process and due to the low percentage of D amino acids in the training set, sequences containing D-residues were excluded. The sequences were further selected following two different strategies.

8.4.7.1 First selection strategy

Since helicity and amphiphilicity often correlate with antimicrobial activity, we selected sequences with a predicted helicity fraction above 0.8 and an Eisenberg hydrophobic moment above 0.3. The thresholds for the predicted helicity fraction and hydrophobic moment were chosen based on the median values of the active sequences in the training and test, respectively 0.83 and 0.31. The filtered sequences were clustered using the RDKit[123] Butina module with a threshold of 10 and the Levenshtein distance as distance function. Sequences containing methionine and sequences with an LD > 5 from the training and test sets of the activity classifier were excluded from all clusters. The center of each cluster was picked, and in addition, one additional compound was selected at random from the clusters containing more than 6 compounds. The workflow resulted in 10 sequences predicted active against gram-negative bacteria and 16 sequences predicted active against gram-positive bacteria 10 sequences for each class were selected for synthesis.

8.4.7.2. Second selection strategy

To avoid the bias that secondary structure evaluation and the clustering might create and to gain a better insight on the activity of the sequences generated by the two transfer learned models, we randomly sampled 20 sequences (10 for each class). four sequences predicted active against gram-positive and five against gram-negative were manually selected. Non-containing methionine sequences with higher distances from the training and test sets of the activity classifier were preferred.

8.4.8. Evaluation metrics

ROC AUC is the area under the ROC curve, and the ROC curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR):

$$TPR = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{TP + FP}$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives predicted by the classifier.

The F1 score is defined as the harmonic mean of precision and recall:

$$Precision = TPR$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

The balanced accuracy is defined as:

$$Balanced\ accuracy = \frac{TPR + \frac{TN}{TN + FN}}{2}$$

The Matthews correlation coefficient (MCC) is a correlation between the observed and the predicted class and it is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

8.4.9 Peptide synthesis

Peptides were synthesized using standard 9-fluorenylmethoxycarbonyl (Fmoc) Solid Phase Peptide Synthesis. All syntheses were performed at 60°C under nitrogen bubbling. 400 mg

Rink Amide AM resin LL (0.26 mmol/g) were used for each peptide. The resin was firstly deprotected twice one minute and four minutes using a deprotection cocktail containing 5% w/v piperazine, 2% v/v 1,8-Diazabicyclo(5.4.0)undéc-7-ene (DBU) and 10% v/v 2-Butanol in N,N-dimethylformamide (DMF). For each amino acid, a doubling coupling was performed (twice eight minutes) using for each coupling 3 mL of 0.2 M of the corresponding Fmoc protected amino acid in DMF, 1.5 mL of 0.5M Oxyma in DMF, and 2 mL of 0.5 M N,N'-diisopropylcarbodiimide (DIC) in DMF. Deprotection steps (double deprotection, one minute, and four minutes) were achieved using the same cocktail described above, except for sequences containing aspartic acid for which a solution of 20% v/v piperidine + 0.7% v/v formic acid in DMF was used to avoid aspartimide and side products formation.

After the last deprotection, peptides were cleaved from the resin using 7 mL of a mixture trifluoroacetic acid/triisopropylsilane/mQ water (TFA/TIS/H_2O) with the corresponding ratios 94/5/1 during three hours. Peptides were then precipitated using approximatively 25 mL of cold terbutylmethyl ether and centrifuged 10 minutes at 4400 rpm. Supernatant was removed and peptides were washed twice with 15 mL of cold terbutylmethyl ether before lyophilization.

8.4.10 Minimal inhibitory concentration

Antimicrobial activity was assayed against *P. aeruginosa* PAO1 (WT), *Acinetobacter baumannii* (ATCC 19606), *K. pneumoniae* (NCTC 418), Methicillin-resistant *Staphylococcus aureus* (COL). To determine the minimal inhibitory concentration (MIC), the broth microdilution method was used. A colony of bacteria was grown in LB (Lysogeny broth) medium overnight at 37 °C. The samples were prepared as stock solutions of 8 mg/mL in H_2O, diluted to the initial concentration of 64 or 128 µg/mL in 300 µL Mueller-Hinton (MH) medium, added to the first well of 96-well microtiter plate (TPP, untreated), and diluted serially

by ½. The concentration of the bacteria was quantified by measuring absorbance at 600 nm and diluted to OD_{600} = 0.022 in MH medium. The sample solutions (150 µL) were mixed with 4 µL diluted bacterial suspension with a final inoculation of about of 5 x 10^5 CFU. The plates were incubated at 37 °C until satisfactory growth (~18 h). For each test, two columns of the plate were kept for sterility control (broth only) and growth control (broth with bacterial inoculums, no antibiotics). The MIC was defined as the lowest concentration of the peptide dendrimer that inhibited visible growth of the tested bacteria, as detected after treatment with MTT.

8.4.11 Hemolysis Assay

Compounds were subjected to a hemolysis assay to assess the hemolytic effect on human red blood cells (hRBCs). The blood was obtained from Interregionale Blutspende SRK AG, Bern, Switzerland. 1.5 mL of whole blood was centrifuged at 3000 rpm for 15 minutes at 4 °C. The plasma was discarded, and the hRBC pellet was re-suspended in 5 mL of PBS (pH 7.4) then centrifuged at 3000 rpm for 5 minutes at 4 °C. The washing of hRBC was repeated three times and the remaining pellet was re-suspended in 10 mL of PBS.

The samples were prepared as the initial concentration of 4000 µg/mL in PBS, added to the first well of 96-well microtiter plate (TPP, untreated) and diluted serially by ½. After diluted, 100 µL of sample was in each well and the final sample concentration was 4000 µg/mL, 2000 µg/mL, 1000 µg/mL, 500 µg/mL, 250 µg/mL, 125 µg/mL, 62.5 µg/mL and 31.3 µg/mL. Controls on each plate included a blank medium control (PBS 100 µL) and a hemolytic activity control (0.1% Triton X-100). 100 µL of hRBC suspension was incubated with 100 µL of each sample in PBS in 96-well plate (Nunc 96-Well Polystyrene Conical Bottom MicroWell Plates). After the plates were incubated for 4 h at room temperature, minimal hemolytic concentration

(MHC) was determined by visual inspection of the wells. 100 µL supernatants was carefully pipetted to a flat bottom, clear wells plate (TPP® tissue culture plates, polystyrene).

8.4.12 Circular dichroism spectroscopy

CD spectra were recorded using a Jasco J-715 spectrometer equipped with a PFD-350S temperature controller and a PS-150J power supply. All experiments were measured using a Hellma Suprasil R 100QS 0.1 cm cuvette. Stock solution (1.00 mg/mL) of dendrimers were freshly prepared in 10 mM phosphate buffer (pH 7.4). For the measurement, the peptides were diluted to 100 µg/mL with buffer and 5 mM Dodecylphosphocholine (DPC, Avanti Polar Lipids, Inc., USA) was added when specified. The range of measurement was 185-260 nm, scan rate was 20 nm/min, pitch 0.5 nm, response 16 sec. and band 1.0 nm. The nitrogen flow was kept above 10 L/min. The blank was recorded under the same conditions and subtracted manually. Each sample was subjected to two accumulations. The cuvettes were washed with 1M HCl, mQ-H_2O and buffer before each measurement. Percentage of different secondary structure was calculated by DichroWeb.

8.4.13 Transmission electron microscopy

Exponential phase of *Pseudomonas aeruginosa* PAO1 and *A. baumannii* were washed with MH medium and treated with **GN1** at the concentration of 10 x MIC. After 2h incubation, 1 ml of the bacteria (OD_{600} = 1) were centrifuged at 12 000 rpm for 3 min and fixed overnight with 2.5% glutaraldehyde (Agar Scientific, Stansted, Essex, UK) in 0.15M HEPES (Fluka, Buchs, Switzerland) with an osmolarity of 670 mOsm and adjusted to a pH of 7.35. The next day, PAO1 were washed with 0.15 M HEPES three times for 5 min, postfixed with 1% OsO4 (SPI Supplies, West Chester, USA) in 0.1 M Na-cacodylate-buffer (Merck, Darmstadt, Germany) at 4°C for 1 h. Thereafter, bacteria cells were washed in 0.1 M Na-cacodylate-buffer three times for 5 min and dehydrated in 70, 80, and 96% ethanol (Alcosuisse, Switzerland) for

15 min each at room temperature. Subsequently, they were immersed in 100% ethanol (Merck, Darmstadt, Germany) three times for 10 min, in acetone (Merck, Darmstadt, Germany) two times for 10 min, and finally in acetone-Epon (1:1) overnight at room temperature. The next day, bacteria cells were embedded in Epon (Fluka, Buchs, Switzerland) and hardened at 60°C for 5 days.

Sections were produced with an ultramicrotome UC6 (Leica Microsystems, Vienna, Austria), first semithin sections (1um) for light microscopy which were stained with a solution of 0.5% toluidine blue O (Merck, Darmstadt, Germany) and then ultrathin sections (70-80 nm) for electron microscopy. The sections, mounted on single-slot copper grids, were stained with 1% uranyl acetate at 40°C for 30 min and 3% lead citrate at RT for 20 min or UranyLess (Electron Microscopy Sciences, Hatfield, UK) at 40°C for 10 min and 3% lead citrate at 25°C for 10 min with an ultrostainer (Leica Microsystems, Vienna, Austria). Sections were then examined with a Tecnai Spirit transmission electron microscope equipped with two digital cameras (Olympus-SIS Veleta CCD Camera, FEI Eagle CCD Camera).

8.4.14 Molecular Dynamics

Molecular dynamics (MD) simulations were performed for the peptides GN1, GP1 and GN2 using GROMACS software version 2018.1 and the gromos53a6 force field[354,361]. The starting topologies were built from the existing X-ray structures. A dodecahedral box was created around the peptide 1.0 nm from the edge of the peptide and filled with extended simple point charge water molecules. Sodium and chloride ions were added to produce an electroneutral solution at a final concentration of 0.15 M NaCl. The energy was minimized using a steepest gradient method to remove any close contacts before the system was subjected to a two-phase position-restrained MD equilibration procedure. The system was first allowed to evolve for 100 ps in a canonical NVT (N is the number of particles, V the system volume, and T the

temperature) ensemble at 300 K before pressure coupling was switched on and the system was equilibrated for an additional 100 ps in the NPT (P is the system pressure) ensemble at 1.0 bar.

8.4.14.1 MD in presence of DPC micelle

MD simulations in the presence of a DPC (n-dodecylphosphocholine) micelle were performed as follows. Parameters and references for the DPC molecule[362] for the GROMOS53a6 forcefield are given in the SI (Note 4). Peptides were manually placed at a distance from the pre-equilibrated micelle (of 65 DPC molecules) equal to the diameter of said peptide. Box, solvation and NVT equilibration procedures were performed as explained previously. For each peptide/micelle system, 10 runs of 50 ns were generated to show the possibility for the peptide to either interact or diffuse away from the micelle. Then, runs of interest were extended up to 250 ns.

8.4.14.3 Clustering of stable structures

To obtain a representative conformer for each SA-MD run, the last 100 ns (10001 frames) of each run were clustered using an RMSD cut-off adapted to get a good balance between the number of clusters and the size of the main cluster. A large number of clusters combined with a very large percentage of structures in the top cluster is an indication of the stability of the one main conformer in each case. The PyMol Molecular Graphics System, version 1.8 (Schrödinger, LLC), was used to create structural models.

Chapter nine – General Conclusions and Outlook

In this thesis, cheminformatics tools for larger molecules and peptides were developed and applied. Two new molecular fingerprints were introduced: MXFP and MAP4. These allowed for a previously unexplored examination of larger molecules databases. The chemical space of molecules that do not obey Lipinski rules within known databases, the natural products space, the human metabolome space, and the virtual and the known chemical space of peptides were encoded and visualized. The visualization was carried out using scatter plots of the first principal components of a PCA of the MXFP feature space, MXFP similarity maps, and MAP4 TMAPs. The resulting maps are all interactive and available to be navigated using different properties.

Furthermore, classical machine learning and the MAP4 fingerprint were used to characterize natural products by their origin, and a genetic algorithm (PDGA) and deep learning were used to generate novel peptide sequences. In addition to the reported studies, PDGA and the deep learning workflow described in chapters seven and eight have been used in additional and yet unpublished projects. For instance, PDGA was used to search analogs of dendritic peptides, and a version of PDGA that uses the RDKit AP in combination with the hemolysis classifier was used to generate anticancer peptides. These projects are carried out in collaboration with two brilliant peptide chemists in the Reymond group: Xingguang Cai and Elena Zakharova. Further development of this generative approach could see a SMILES-based adaptation to explore the chemical space of non-peptidic large molecules.

An important topic that has been only partially addressed in this thesis is the role of stereochemistry and conformation in peptide activity. In fact, MXFP and MAP4 are 2D fingerprints, and they do not explicitly encode stereochemistry.[14,18] On the one hand, the correlation between the conformation of a peptide and its properties is well accepted.[333,363] For instance, most active peptides in DBAASP[332] were predicted to be amphiphilic helixes, an observation that led us to use the same estimation to guide the selection of peptide sequences in output of the RNN generative workflow.[364] Furthermore, in a recent study,[37] Gao *et al.* compared 2D-fingerprint-based machine learning models with the state-of-the-art 3D structure-based models in different drug discovery-related applications. They showed that 2D fingerprint-based models had the same capabilities as the state-of-the-art 3D approach in different drug discovery-related tasks, except estimating protein-ligand binding affinity based on ligand and protein information. These findings highlight the possible limits of 2D models in handling complex systems that can assume multiple conformations.

On the other hand, in a recent publication, Siriwardena *et al.* showed that some peptides that were widely believed to require folding for activity also work when stereorandomized.[365] Furthermore, Senior *et al.* demonstrated that sequence information alone could be used to predict a protein structure with machine learning even when only a few homologs are available.[366] Taken together, these factors highlight the validity of a fast 2D fingerprints-guided approach to lead early findings and its orthogonality to a three-dimensional analysis that considers stereochemistry and conformation for further optimization processes.

A possible future direction of the work presented in this thesis could be investigating the space of another important class of macromolecules: glycans. Glycans are generally reported with specific nomenclatures. The first example of glycan-specific nomenclatures is the Kornfeld model, where monosaccharides are represented by geometric symbols and glycosidic linkages by a simple line.[367] Since then, the model has been expanded, and the

current version, called Symbol Nomenclature for Glycans (SNFG), includes colors and new monosaccharides.[368] SNFG is a symbolic graphic, and its application within fingerprint calculation, visualization, and structure generations is not trivial. However, SNFG can be converted into Web3 Unique Representation of Carbohydrate Structure (WURST),[369] a unique string representation of glycan structures partially based on SMILES. WURST itself could be used as input for a machine learning workflow or converted into SMILES and encoded with molecular fingerprints. This latter approach would have the advantage of also allowing for the encoding of glycosylated molecules such as glycosylated peptides and lipids.

References

(1) Wishart, D. S. Introduction to Cheminformatics. *Curr. Protoc. Bioinforma.* **2007**, *18* (1), 14.1.1-14.1.9. https://doi.org/10.1002/0471250953.bi1401s18.

(2) Levine, D. P. Vancomycin: A History. *Clin. Infect. Dis.* **2006**, *42* (Supplement_1), S5–S12. https://doi.org/10.1086/491709.

(3) Pires, J.; Siriwardena, T. N.; Stach, M.; Tinguely, R.; Kasraian, S.; Luzzaro, F.; Leib, S. L.; Darbre, T.; Reymond, J.-L.; Endimiani, A. In Vitro Activity of the Novel Antimicrobial Peptide Dendrimer G3KL against Multidrug-Resistant Acinetobacter Baumannii and Pseudomonas Aeruginosa. *Antimicrob. Agents Chemother.* **2015**, *59* (12), 7915–7918. https://doi.org/10.1128/AAC.01853-15.

(4) Siriwardena, T. N.; Capecchi, A.; Gan, B.-H.; Jin, X.; He, R.; Wei, D.; Ma, L.; Köhler, T.; van Delden, C.; Javor, S.; Reymond, J.-L. Optimizing Antimicrobial Peptide Dendrimers in Chemical Space. *Angew. Chem. Int. Ed.* **2018**, *57* (28), 8483–8487. https://doi.org/10.1002/anie.201802837.

(5) Bonaventura, I. D.; Baeriswyl, S.; Capecchi, A.; Gan, B.-H.; Jin, X.; N. Siriwardena, T.; He, R.; Köhler, T.; Pompilio, A.; Bonaventura, G. D.; Delden, C.; Javor, S.; Reymond, J.-L. An Antimicrobial Bicyclic Peptide from Chemical Space against Multidrug Resistant Gram-Negative Bacteria. *Chem. Commun.* **2018**, *54* (40), 5130–5133. https://doi.org/10.1039/C8CC02412J.

(6) Bonaventura, I. D.; Jin, X.; Visini, R.; Probst, D.; Javor, S.; Gan, B.-H.; Michaud, G.; Natalello, A.; Maria Doglia, S.; Köhler, T.; Delden, C.; Stocker, A.; Darbre, T.; Reymond, J.-L. Chemical Space Guided Discovery of Antimicrobial Bridged Bicyclic Peptides against Pseudomonas Aeruginosa and Its Biofilms. *Chem. Sci.* **2017**, *8* (10), 6784–6798. https://doi.org/10.1039/C7SC01314K.

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

(8) Probst, D.; Reymond, J.-L. A Probabilistic Molecular Fingerprint for Big Data Settings. *J. Cheminformatics* **2018**, *10* (1), 66. https://doi.org/10.1186/s13321-018-0321-8.

(9) 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.

(10) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J Chem Inf Comput Sci* **1985**, *25* (2), 64–73. https://doi.org/10.1021/ci00046a002.

(11) Awale, M.; Reymond, J.-L. Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2015**, *55* (8), 1509–1516. https://doi.org/10.1021/acs.jcim.5b00182.

(12) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminformatics* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.

(13) 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), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(14) Capecchi, A.; Awale, M.; Probst, D.; Reymond, J.-L. PubChem and ChEMBL beyond Lipinski. *Mol. Inform.* **2019**, *38* (5). https://doi.org/10.1002/minf.201900016.

(15) 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.

(16) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminformatics* **2013**, *5* (1), 26. https://doi.org/10.1186/1758-2946-5-26.
(17) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608–D617. https://doi.org/10.1093/nar/gkx1089.
(18) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.
(19) van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent. Sci.* **2019**, *5* (11), 1824–1833. https://doi.org/10.1021/acscentsci.9b00806.
(20) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminformatics* **2021**, *13* (1), 2. https://doi.org/10.1186/s13321-020-00478-9.
(21) Capecchi, A.; Reymond, J.-L. Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning. *Biomolecules* **2020**, *10* (10), 1385. https://doi.org/10.3390/biom10101385.
(22) Capecchi, A.; Reymond, J.-L. Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database. **2021**. https://doi.org/10.33774/chemrxiv-2021-gxjgc.
(23) Capecchi, A.; Reymond, J.-L. Peptides in Chemical Space. *Med. Drug Discov.* **2021**, *9*, 100081. https://doi.org/10.1016/j.medidd.2021.100081.
(24) Probst, D.; Reymond, J.-L.; Wren, J. FUn: A Framework for Interactive Visualizations of Large, High-Dimensional Datasets on the Web. *Bioinformatics* **2018**, *34* (8), 1433–1435. https://doi.org/10.1093/bioinformatics/btx760.
(25) Capecchi, A.; Zhang, A.; Reymond, J.-L. Populating Chemical Space with Peptides Using a Genetic Algorithm. *J. Chem. Inf. Model.* **2020**, *60* (1), 121–132. https://doi.org/10.1021/acs.jcim.9b01014.
(26) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; Delden, C. van; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem. Sci.* **2021**, *12* (26), 9221–9232. https://doi.org/10.1039/D1SC01713F.
(27) Hansch, Corwin.; Fujita, Toshio. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616–1626. https://doi.org/10.1021/ja01062a035.
(28) McCulloch, W. S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* **1943**, *5* (4), 115–133. https://doi.org/10.1007/BF02478259.
(29) Brown, F. K. Chapter 35 - Chemoinformatics: What Is It and How Does It Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*; Bristol, J. A., Ed.; Academic Press, 1998; Vol. 33, pp 375–384. https://doi.org/10.1016/S0065-7743(08)61100-8.
(30) Pereira, D. A.; Williams, J. A. Origin and Evolution of High Throughput Screening. *Br. J. Pharmacol.* **2007**, *152* (1), 53–61. https://doi.org/10.1038/sj.bjp.0707373.
(31) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(32) Martinez-Mayorga, K.; Madariaga-Mazon, A.; Medina-Franco, J. L.; Maggiora, G. The Impact of Chemoinformatics on Drug Discovery in the Pharmaceutical Industry. *Expert Opin. Drug Discov.* **2020**, *15* (3), 293–306. https://doi.org/10.1080/17460441.2020.1696307.
(33) Guha, R.; Willighagen, E.; Zdrazil, B.; Jeliazkova, N. What Is the Role of Cheminformatics in a Pandemic? *J. Cheminformatics* **2021**, *13* (1), 16. https://doi.org/10.1186/s13321-021-00491-6.
(34) Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Mol. J. Synth. Chem. Nat. Prod. Chem.* **2002**, *7* (8), 566–600. https://doi.org/10.3390/70800566.
(35) Awale, M.; Riniker, S.; Kramer, C. Matched Molecular Series Analysis for ADME Property Prediction. *J. Chem. Inf. Model.* **2020**, *60* (6), 2903–2914. https://doi.org/10.1021/acs.jcim.0c00269.
(36) Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.* **2012**, *52* (5), 1103–1113. https://doi.org/10.1021/ci300030u.
(37) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D Fingerprints Still Valuable for Drug Discovery? *Phys. Chem. Chem. Phys.* **2020**, *22* (16), 8373–8390. https://doi.org/10.1039/D0CP00305K.
(38) 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 Investig. Drugs J.* **2006**, *9* (3), 199–204.
(39) 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.
(40) Dang, Q. H. *Secure Hash Standard*; 180–4; National Institute of Standards and Technology, 2015. https://doi.org/10.6028/NIST.FIPS.180-4.
(41) Broder, A. Z. On the Resemblance and Containment of Documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*; 1997; pp 21–29. https://doi.org/10.1109/SEQUEN.1997.666900.
(42) Bawa, M.; Condie, T.; Ganesan, P. LSH Forest: Self-Tuning Indexes for Similarity Search. In *Proceedings of the 14th international conference on World Wide Web*; WWW '05; Association for Computing Machinery: Chiba, Japan, 2005; pp 651–660. https://doi.org/10.1145/1060745.1060840.
(43) Reutlinger, M.; Koch, C. P.; Reker, D.; Todoroff, N.; Schneider, P.; Rodrigues, T.; Schneider, G. Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Mol. Inform.* **2013**, *32* (2), 133–138. https://doi.org/10.1002/minf.201200141.
(44) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801. https://doi.org/10.1021/jm0608356.
(45) Awale, M.; Visini, R.; Probst, D.; Arús-Pous, J.; Reymond, J.-L. Chemical Space: Big Data Challenge for Molecular Diversity. *Chim. Int. J. Chem.* **2017**, *71* (10), 661–666. https://doi.org/10.2533/chimia.2017.661.
(46) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730. https://doi.org/10.1021/ar500432k.
(47) Reymond, J.-L.; Deursen, R. van; C. Blum, L.; Ruddigkeit, L. Chemical Space as a Source for New Drugs. *MedChemComm* **2010**, *1* (1), 30–38. https://doi.org/10.1039/C0MD00020E.
(48) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16* (1), 3–50. https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.
(49) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.
(50) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 374–380. https://doi.org/10.1021/ci0255782.

(51) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342–353. https://doi.org/10.1021/ci600423u.
(52) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131* (25), 8732–8733. https://doi.org/10.1021/ja902302h.
(53) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.
(54) Li, X.; Xu, Y.; Yao, H.; Lin, K. Chemical Space Exploration Based on Recurrent Neural Networks: Applications in Discovering Kinase Inhibitors. *J. Cheminformatics* **2020**, *12* (1), 42. https://doi.org/10.1186/s13321-020-00446-3.
(55) Öztürk, H.; Özgür, A.; Schwaller, P.; Laino, T.; Ozkirimli, E. Exploring Chemical Space Using Natural Language Processing Methodologies for Drug Discovery. *Drug Discov. Today* **2020**, *25* (4), 689–705. https://doi.org/10.1016/j.drudis.2020.01.020.
(56) Hinton, G.; Roweis, S. Stochastic Neighbor Embedding. *Neural Information Processing Systems* **2002**, 8.
(57) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* **2018**.
(58) F.R.S, K. P. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2* (11), 559–572. https://doi.org/10.1080/14786440109462720.
(59) Jolliffe, I. T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2016**, *374* (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.
(60) Kruskal, J. B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Am. Math. Soc.* **1956**, *7* (1), 48–50. https://doi.org/10.1090/S0002-9939-1956-0078686-7.
(61) Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *J. Chem. Inf. Model.* **2018**, *58* (8), 1518–1532. https://doi.org/10.1021/acs.jcim.8b00302.
(62) Larsson, J.; Gottfries, J.; Bohlin, L.; Backlund, A. Expanding the ChemGPS Chemical Space with Natural Products. *J. Nat. Prod.* **2005**, *68* (7), 985–991. https://doi.org/10.1021/np049655u.
(63) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci.* **2005**, *102* (48), 17272–17277.
(64) Lachance, H.; Wetzel, S.; Kumar, K.; Waldmann, H. Charting, Navigating, and Populating Natural Product Chemical Space for Drug Discovery. *J. Med. Chem.* **2012**, *55* (13), 5989–6001. https://doi.org/10.1021/jm300288g.
(65) Rosén, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel Chemical Space Exploration via Natural Products. *J. Med. Chem.* **2009**, *52* (7), 1953–1962. https://doi.org/10.1021/jm801514w.
(66) Tao, L.; Zhu, F.; Qin, C.; Zhang, C.; Chen, S.; Zhang, P.; Zhang, C.; Tan, C.; Gao, C.; Chen, Z.; Jiang, Y.; Chen, Y. Z. Clustered Distribution of Natural Product Leads of Drugs in the Chemical Space as Influenced by the Privileged Target-Sites. *Sci. Rep.* **2015**, *5* (1), 9325. https://doi.org/10.1038/srep09325.
(67) Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: A New Look at the Natural Product Chemical Space. *Mol. Inform. n/a* (n/a). https://doi.org/10.1002/minf.202100068.
(68) Díaz-Eufracio, B. I.; Palomino-Hernández, O.; Houghten, R. A.; Medina-Franco, J. L. Exploring the Chemical Space of Peptides for Drug Discovery: A Focus on Linear and Cyclic

Penta-Peptides. *Mol. Divers.* **2018**, *22* (2), 259–267. https://doi.org/10.1007/s11030-018-9812-9.

(69) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(70) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (Database issue), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(71) Valeur, E.; Guéret, S. M.; Adihou, H.; Gopalakrishnan, R.; Lemurell, M.; Waldmann, H.; Grossmann, T. N.; Plowright, A. T. New Modalities for Challenging Targets in Drug Discovery. *Angew. Chem. Int. Ed.* **2017**, *56* (35), 10294–10323. https://doi.org/10.1002/anie.201611914.

(72) Rigden, D. J.; Fernández, X. M. The 2018 Nucleic Acids Research Database Issue and the Online Molecular Biology Database Collection. *Nucleic Acids Res.* **2018**, *46* (D1), D1–D7. https://doi.org/10.1093/nar/gkx1235.

(73) Shtatland, T.; Guettler, D.; Kossodo, M.; Pivovarov, M.; Weissleder, R. PepBank - a Database of Peptides Based on Sequence Text Mining and Public Peptide Data Sources. *BMC Bioinformatics* **2007**, *8* (1), 280. https://doi.org/10.1186/1471-2105-8-280.

(74) Wang, J.; Yin, T.; Xiao, X.; He, D.; Xue, Z.; Jiang, X.; Wang, Y. StraPep: A Structure Database of Bioactive Peptides. *Database* **2018**, *2018* (bay038). https://doi.org/10.1093/database/bay038.

(75) Newburger, D. E.; Natsoulis, G.; Grimes, S.; Bell, J. M.; Davis, R. W.; Batzoglou, S.; Ji, H. P. The Human OligoGenome Resource: A Database of Oligonucleotide Capture Probes for Resequencing Target Regions across the Human Genome. *Nucleic Acids Res.* **2012**, *40* (D1), D1137–D1143. https://doi.org/10.1093/nar/gkr973.

(76) Fahy, E.; Subramaniam, S.; Brown, H. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Seyama, Y.; Shaw, W.; Shimizu, T.; Spener, F.; van Meer, G.; VanNieuwenhze, M. S.; White, S. H.; Witztum, J. L.; Dennis, E. A. A Comprehensive Classification System for Lipids. *J. Lipid Res.* **2005**, *46* (5), 839–861. https://doi.org/10.1194/jlr.E400004-JLR200.

(77) Kuo, T.-C.; Tseng, Y. J. LipidPedia: A Comprehensive Lipid Knowledgebase. *Bioinforma. Oxf. Engl.* **2018**, *34* (17), 2982–2987. https://doi.org/10.1093/bioinformatics/bty213.

(78) Campbell, M. P.; Peterson, R.; Mariethoz, J.; Gasteiger, E.; Akune, Y.; Aoki-Kinoshita, K. F.; Lisacek, F.; Packer, N. H. UniCarbKB: Building a Knowledge Platform for Glycoproteomics. *Nucleic Acids Res.* **2014**, *42* (Database issue), D215–D221. https://doi.org/10.1093/nar/gkt1128.

(79) Birch, J.; Van Calsteren, M.-R.; Pérez, S.; Svensson, B. The Exopolysaccharide Properties and Structures Database: EPS-DB. Application to Bacterial Exopolysaccharides. *Carbohydr. Polym.* **2019**, *205*, 565–570. https://doi.org/10.1016/j.carbpol.2018.10.063.

(80) Clerc, O.; Mariethoz, J.; Rivet, A.; Lisacek, F.; Pérez, S.; Ricard-Blum, S. A Pipeline to Translate Glycosaminoglycan Sequences into 3D Models. Application to the Exploration of Glycosaminoglycan Conformational Space. *Glycobiology* **2019**, *29* (1), 36–44. https://doi.org/10.1093/glycob/cwy084.

(81) Berhanu, W. M.; Ibrahim, M. A.; Pillai, G. G.; Oliferenko, A. A.; Khelashvili, L.; Jabeen, F.; Mirza, B.; Ansari, F. L.; ul-Haq, I.; El-Feky, S. A.; Katritzky, A. R. Similarity Analysis, Synthesis, and Bioassay of Antibacterial Cyclic Peptidomimetics. *Beilstein J. Org. Chem.* **2012**, *8* (1), 1146–1160. https://doi.org/10.3762/bjoc.8.128.

(82) Santos, G. B.; Ganesan, A.; Emery, F. S. Oral Administration of Peptide-Based Drugs: Beyond Lipinski's Rule. *ChemMedChem* **2016**, *11* (20), 2245–2251. https://doi.org/10.1002/cmdc.201600288.

(83) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.
(84) Poongavanam, V.; Doak, B. C.; Kihlberg, J. Opportunities and Guidelines for Discovery of Orally Absorbed Drugs in beyond Rule of 5 Space. *Curr. Opin. Chem. Biol.* **2018**, *44*, 23–29. https://doi.org/10.1016/j.cbpa.2018.05.010.
(85) DeGoey, D. A.; Chen, H.-J.; Cox, P. B.; Wendt, M. D. Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection. *J. Med. Chem.* **2018**, *61* (7), 2636–2651. https://doi.org/10.1021/acs.jmedchem.7b00717.
(86) Leeson, P. D. Molecular Inflation, Attrition and the Rule of Five. *Adv. Drug Deliv. Rev.* **2016**, *101*, 22–33. https://doi.org/10.1016/j.addr.2016.01.018.
(87) Awale, M.; van Deursen, R.; Reymond, J.-L. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013**, *53* (2), 509–518. https://doi.org/10.1021/ci300513m.
(88) Awale, M.; Reymond, J.-L. A Multi-Fingerprint Browser for the ZINC Database. *Nucleic Acids Res.* **2014**, *42* (W1), W234–W239. https://doi.org/10.1093/nar/gku379.
(89) Awale, M.; Reymond, J.-L. Web-Based 3D-Visualization of the DrugBank Chemical Space. *J. Cheminformatics* **2016**, *8* (1), 25. https://doi.org/10.1186/s13321-016-0138-2.
(90) Awale, M.; Reymond, J.-L. The Polypharmacology Browser: A Web-Based Multi-Fingerprint Target Prediction Tool Using ChEMBL Bioactivity Data. *J. Cheminformatics* **2017**, *9* (1), 11. https://doi.org/10.1186/s13321-017-0199-x.
(91) Awale, M.; Probst, D.; Reymond, J.-L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2017**, *57* (4), 643–649. https://doi.org/10.1021/acs.jcim.6b00690.
(92) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4* (11), 1803–1805. https://doi.org/10.1002/cmdc.200900317.
(93) 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.
(94) Deursen, R. van; Blum, L. C.; Reymond, J.-L. A Searchable Map of PubChem. *J. Chem. Inf. Model.* **2010**, *50* (11), 1924–1934. https://doi.org/10.1021/ci100237q.
(95) Blum, L. C.; van Deursen, R.; Reymond, J.-L. Visualisation and Subsets of the Chemical Universe Database GDB-13 for Virtual Screening. *J. Comput. Aided Mol. Des.* **2011**, *25* (7), 637–647. https://doi.org/10.1007/s10822-011-9436-y.
(96) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 128–136. https://doi.org/10.1021/ci950275b.
(97) Awale, M.; Jin, X.; Reymond, J.-L. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminformatics* **2015**, *7* (1), 3. https://doi.org/10.1186/s13321-014-0051-5.
(98) Jin, X.; Awale, M.; Zasso, M.; Kostro, D.; Patiny, L.; Reymond, J.-L. PDB-Explorer: A Web-Based Interactive Map of the Protein Data Bank in Shape Space. *BMC Bioinformatics* **2015**, *16*, 339. https://doi.org/10.1186/s12859-015-0776-9.
(99) Probst, D.; Reymond, J.-L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J. Chem. Inf. Model.* **2018**, *58* (1), 1–7. https://doi.org/10.1021/acs.jcim.7b00425.
(100) Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 987–1003. https://doi.org/10.1021/ci025599w.
(101) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70* (5), 393–412. https://doi.org/10.1111/j.1747-0285.2007.00579.x.

(102) 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.
(103) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015**, *55* (1), 84–94. https://doi.org/10.1021/ci500575y.
(104) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. https://doi.org/10.1021/ci500588j.
(105) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3* (2), 157–166. https://doi.org/10.1021/cc0000388.
(106) Delalande, C.; Awale, M.; Rubin, M.; Probst, D.; Ozhathil, L. C.; Gertsch, J.; Abriel, H.; Reymond, J.-L. Optimizing TRPM4 Inhibitors in the MHFP6 Chemical Space. *Eur. J. Med. Chem.* **2019**, *166*, 167–177. https://doi.org/10.1016/j.ejmech.2019.01.048.
(107) Visini, R.; Arús-Pous, J.; Awale, M.; Reymond, J.-L. Virtual Exploration of the Ring Systems Chemical Universe. *J. Chem. Inf. Model.* **2017**, *57* (11), 2707–2718. https://doi.org/10.1021/acs.jcim.7b00457.
(108) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (Database issue), D1202–D1213. https://doi.org/10.1093/nar/gkv951.
(109) Yonchev, D.; Dimova, D.; Stumpfe, D.; Vogt, M.; Bajorath, J. Redundancy in Two Major Compound Databases. *Drug Discov. Today* **2018**, *23* (6), 1183–1186. https://doi.org/10.1016/j.drudis.2018.03.005.
(110) Bon, I.; Lembo, D.; Rusnati, M.; Clò, A.; Morini, S.; Miserocchi, A.; Bugatti, A.; Grigolon, S.; Musumeci, G.; Landolfo, S.; Re, M. C.; Gibellini, D. Peptide-Derivatized SB105-A10 Dendrimer Inhibits the Infectivity of R5 and X4 HIV-1 Strains in Primary PBMCs and Cervicovaginal Histocultures. *PLOS ONE* **2013**, *8* (10), e76482. https://doi.org/10.1371/journal.pone.0076482.
(111) Egbert, M.; Whitty, A.; Keserű, G. M.; Vajda, S. Why Some Targets Benefit from beyond Rule of Five Drugs. *J. Med. Chem.* **2019**, *62* (22), 10005–10025. https://doi.org/10.1021/acs.jmedchem.8b01732.
(112) Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in Early Drug Discovery: Focus on the beyond-Rule-of-5 Chemical Space. *Drug Discov. Today* **2020**. https://doi.org/10.1016/j.drudis.2020.01.012.
(113) J. Maple, H.; Clayden, N.; Baron, A.; Stacey, C.; Felix, R. Developing Degraders: Principles and Perspectives on Design and Chemical Space. *MedChemComm* **2019**, *10* (10), 1755–1764. https://doi.org/10.1039/C9MD00272C.
(114) Bender, A.; Brown, N. Special Issue: Cheminformatics in Drug Discovery. *ChemMedChem* **2018**, *13* (6), 467–469. https://doi.org/10.1002/cmdc.201800123.
(115) Bajusz, D.; Rácz, A.; Héberger, K. 3.14 - Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In *Comprehensive Medicinal Chemistry III*; Chackalamannil, S., Rotella, D., Ward, S. E., Eds.; Elsevier: Oxford, 2017; pp 329–378. https://doi.org/10.1016/B978-0-12-409547-2.12345-5.
(116) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11* (23–24), 1046–1053.
(117) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J Chem Inf Model* **2012**, *52* (4), 867–881. https://doi.org/10.1021/ci200528d.
(118) Naveja, J. J.; Medina-Franco, J. L. ChemMaps: Towards an Approach for Visualizing the Chemical Space Based on Adaptive Satellite Compounds. *F1000Res* **2017**, *6*, Chem Inf Sci-1134. https://doi.org/10.12688/f1000research.12095.2.
(119) Awale, M.; Reymond, J. L. Web-Based Tools for Polypharmacology Prediction. *Methods Mol Biol* **2019**, *1888*, 255–272. https://doi.org/10.1007/978-1-4939-8891-4_15.

(120) Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J Chem Inf Model* **2019**, *59* (1), 10–17. https://doi.org/10.1021/acs.jcim.8b00524.
(121) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew Chem Int Ed Engl* **1999**, 38 (19), 2894–2896.
(122) Siriwardena, T. N.; Lüscher, A.; Köhler, T.; van Delden, C.; Javor, S.; Reymond, J.-L. Antimicrobial Peptide Dendrimer Chimera. *Helv. Chim. Acta* **2019**, *102* (4), e1900034. https://doi.org/10.1002/hlca.201900034.
(123) RDKit https://www.rdkit.org/ (accessed 2020 -06 -02).
(124) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
(125) Gionis, A.; Indyk, P.; Motwani, R. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*; VLDB '99; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp 518–529. https://doi.org/10.5555/645925.671516.
(126) Andoni, A.; Razenshteyn, I.; Nosatzki, N. S. LSH Forest: Practical Algorithms Made Theoretical. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*; Proceedings; Society for Industrial and Applied Mathematics, 2017; pp 67–78. https://doi.org/10.1137/1.9781611974782.5.
(127) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037.
(128) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.
(129) Poux, S.; Arighi, C. N.; Magrane, M.; Bateman, A.; Wei, C.-H.; Lu, Z.; Boutet, E.; Bye-A-Jee, H.; Famiglietti, M. L.; Roechert, B.; Consortium, T. U. On Expert Curation and Sustainability: UniProtKB/Swiss-Prot as a Case Study. *bioRxiv* **2016**, 094011. https://doi.org/10.1101/094011.
(130) UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506–D515. https://doi.org/10.1093/nar/gky1049.
(131) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order—An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55* (10), 2111–2120. https://doi.org/10.1021/acs.jcim.5b00543.
(132) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Cheminformatics* **2013**, *5* (1), 24. https://doi.org/10.1186/1758-2946-5-24.
(133) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Jarrod Millman, K.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.; Polat, İ.; Feng, Y.; Moore, E. W.; Vand erPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Contributors, S. 1. 0. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.
(134) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184. https://doi.org/10.1021/ci8002649.
(135) Hollander, M.; Wolfe, D. A.; Chicken, E. *Nonparametric Statistical Methods, Chapter 7, p. 316*; Wiley Series in Probability and Statistics; Wiley, 2013.

(136) Pereira, D. G.; Afonso, A.; Medeiros, F. M. Overview of Friedman's Test and Post-Hoc Analysis. *Commun. Stat. - Simul. Comput.* **2015**, *44* (10), 2636–2653. https://doi.org/10.1080/03610918.2014.931971.
(137) Ozhathil, L. C.; Delalande, C.; Bianchi, B.; Nemeth, G.; Kappel, S.; Thomet, U.; Ross-Kaschitza, D.; Simonin, C.; Rubin, M.; Gertsch, J.; Lochner, M.; Peinelt, C.; Reymond, J. L.; Abriel, H. Identification of Potent and Selective Small Molecule Inhibitors of the Cation Channel TRPM4. *Br J Pharmacol* **2018**, *175* (12), 2504–2519. https://doi.org/10.1111/bph.14220.
(138) Klein, P. N. *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*; Proceedings; Society for Industrial and Applied Mathematics, 2017. https://doi.org/10.1137/1.9781611974782.
(139) Pham, J. V.; Yilma, M. A.; Feliz, A.; Majid, M. T.; Maffetone, N.; Walker, J. R.; Kim, E.; Cho, H. J.; Reynolds, J. M.; Song, M. C.; Park, S. R.; Yoon, Y. J. A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Front. Microbiol.* **2019**, *10*, 1404. https://doi.org/10.3389/fmicb.2019.01404.
(140) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111. https://doi.org/10.1021/acs.jcim.7b00341.
(141) Osada, H.; Nogawa, T. Systematic Isolation of Microbial Metabolites for Natural Products Depository (NPDepo). *Pure Appl. Chem.* **2011**, *84* (6), 1407–1420. https://doi.org/10.1351/PAC-CON-11-08-11.
(142) Grabowski, K.; Baringhaus, K.-H.; Schneider, G. Scaffold Diversity of Natural Products: Inspiration for Combinatorial Library Design. *Nat. Prod. Rep.* **2008**, *25* (5), 892–904. https://doi.org/10.1039/B715668P.
(143) Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J. A.; Tagliabue, S. G.; Todeschini, R.; Schneider, G. Scaffold Hopping from Natural Products to Synthetic Mimetics by Holistic Molecular Similarity. *Commun. Chem.* **2018**, *1* (1), 1–9. https://doi.org/10.1038/s42004-018-0043-x.
(144) Fraser, L.-A.; Mulholland, D. A.; Fraser, D. D. Classification of Limonoids and Protolimonoids Using Neural Networks. *Phytochem. Anal.* **1997**, *8* (6), 301–311. https://doi.org/10.1002/(SICI)1099-1565(199711/12)8:6<301::AID-PCA373>3.0.CO;2-2.
(145) Martínez-Treviño, S. H.; Uc-Cetina, V.; Fernández-Herrera, M. A.; Merino, G. Prediction of Natural Product Classes Using Machine Learning and 13C NMR Spectroscopic Data. *J. Chem. Inf. Model.* **2020**, *7* (69), 3376–3386. https://doi.org/10.1021/acs.jcim.0c00293.
(146) Rupp, M.; Bauer, M. R.; Wilcken, R.; Lange, A.; Reutlinger, M.; Boeckler, F. M.; Schneider, G. Machine Learning Estimates of Natural Product Conformational Energies. *PLOS Comput. Biol.* **2014**, *10* (1), e1003400. https://doi.org/10.1371/journal.pcbi.1003400.
(147) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43. https://doi.org/10.3390/biom9020043.
(148) Rupp, M.; Schroeter, T.; Steri, R.; Zettl, H.; Proschak, E.; Hansen, K.; Rau, O.; Schwarz, O.; Müller-Kuhrt, L.; Schubert-Zsilavecz, M.; Müller, K.-R.; Schneider, G. From Machine Learning to Natural Product Derivatives That Selectively Activate Transcription Factor PPARγ. *ChemMedChem* **2010**, *5* (2), 191–194. https://doi.org/10.1002/cmdc.200900469.
(149) Awale, M.; Sirockin, F.; Stiefl, N.; Reymond, J.-L. Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. *J. Chem. Inf. Model.* **2019**, *59* (4), 1347–1356. https://doi.org/10.1021/acs.jcim.8b00902.
(150) Wang, Y.; Jafari, M.; Tang, Y.; Tang, J. Predicting Meridian in Chinese Traditional Medicine Using Machine Learning Approaches. *PLoS Comput. Biol.* **2019**, *15* (11). https://doi.org/10.1371/journal.pcbi.1007249.
(151) Zhang, R.; Li, X.; Zhang, X.; Qin, H.; Xiao, W. Machine Learning Approaches for Elucidating the Biological Effects of Natural Products. *Nat. Prod. Rep.* **2020**. https://doi.org/10.1039/D0NP00043D.
(152) Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26* (3), 297–302. https://doi.org/10.2307/1932409.

(153) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9* (Nov), 2579–2605.
(154) Broder, A. Z.; Charikar, M.; Frieze, A. M.; Mitzenmacher, M. Min-Wise Independent Permutations. *J. Comput. Syst. Sci.* **1998**, *60*, 327–336.
(155) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868–873. https://doi.org/10.1021/ci990307l.
(156) Shi, C.; Borchardt, T. B. JRgui: A Python Program of Joback and Reid Method. *ACS Omega* **2017**, *2* (12), 8682–8688. https://doi.org/10.1021/acsomega.7b01464.
(157) JOBACK, K. G.; REID, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun.* **1987**, *57* (1–6), 233–243. https://doi.org/10.1080/00986448708960487.
(158) Daylight https://www.daylight.com/ (accessed 2020 -07 -17).
(159) Noble, W. S. What Is a Support Vector Machine? *Nat. Biotechnol.* **2006**, *24* (12), 1565–1567. https://doi.org/10.1038/nbt1206-1565.
(160) Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*; MIT Press, 1999; pp 61–74.
(161) Vert, JP.; Tsuda, K.; Schölkopf, B. A Primer on Kernel Methods. In *Kernel Methods in Computational Biology*; Biologische Kybernetik: Cambridge, MA, USA, 2004; pp 35–70.
(162) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(163) Gallegos, D. A.; Saurí, J.; Cohen, R. D.; Wan, X.; Videau, P.; Vallota-Eastman, A. O.; Shaala, L. A.; Youssef, D. T. A.; Williamson, R. T.; Martin, G. E.; Philmus, B.; Sikora, A. E.; Ishmael, J. E.; McPhail, K. L. Jizanpeptins, Cyanobacterial Protease Inhibitors from a Symploca Sp. Cyanobacterium Collected in the Red Sea. *J. Nat. Prod.* **2018**, *81* (6), 1417–1425. https://doi.org/10.1021/acs.jnatprod.8b00117.
(164) Mao, X.-M.; Xu, W.; Li, D.; Yin, W.-B.; Chooi, Y.-H.; Li, Y.-Q.; Tang, Y.; Hu, Y. Epigenetic Genome Mining of an Endophytic Fungus Leads to the Pleiotropic Biosynthesis of Natural Products. *Angew. Chem. Int. Ed.* **2015**, *54* (26), 7592–7596. https://doi.org/10.1002/anie.201502452.
(165) Dion, H. W.; Woo, P. W. K.; Willmer, N. E.; Kern, D. L.; Onaga, J.; Fusari, S. A. Butirosin, a New Aminoglycosidic Antibiotic Complex: Isolation and Characterization. *Antimicrob. Agents Chemother.* **1972**, *2* (2), 84–88. https://doi.org/10.1128/AAC.2.2.84.
(166) Tatsuda, D.; Momose, I.; Someno, T.; Sawa, R.; Kubota, Y.; Iijima, M.; Kunisada, T.; Watanabe, T.; Shibasaki, M.; Nomoto, A. Quinofuracins A–E, Produced by the Fungus Staphylotrichum Boninense PF1444, Show P53-Dependent Growth Suppression. *J. Nat. Prod.* **2015**, *78* (2), 188–195. https://doi.org/10.1021/np500581m.
(167) Zhang, Y.; Liu, S.; Liu, H.; Liu, X.; Che, Y. Cycloaspeptides F and G, Cyclic Pentapeptides from a Cordyceps-Colonizing Isolate of Isaria Farinosa. *J. Nat. Prod.* **2009**, *72* (7), 1364–1367. https://doi.org/10.1021/np900205m.
(168) Tsuji, N.; Kobayashi, M.; Kamigauchi, T.; Yoshimura, Y.; Terui, Y. New Glycopeptide Antibiotics. I. The Structures of Orienticins. *J. Antibiot. (Tokyo)* **1988**, *41* (6), 819–822. https://doi.org/10.7164/antibiotics.41.819.
(169) Kim, M. C.; Hwang, E.; Kim, T.; Ham, J.; Kim, S. Y.; Kwon, H. C. Nocatriones A and B, Photoprotective Tetracenediones from a Marine-Derived Nocardiopsis Sp. *J. Nat. Prod.* **2014**, *77* (10), 2326–2330. https://doi.org/10.1021/np5006086.
(170) Li, X.-B.; Zhou, Y.-H.; Zhu, R.-X.; Chang, W.-Q.; Yuan, H.-Q.; Gao, W.; Zhang, L.-L.; Zhao, Z.-T.; Lou, H.-X. Identification and Biological Evaluation of Secondary Metabolites from the Endolichenic Fungus Aspergillus Versicolor. *Chem. Biodivers.* **2015**, *12* (4), 575–592. https://doi.org/10.1002/cbdv.201400146.
(171) Spyere, A.; Rowley, D. C.; Jensen, P. R.; Fenical, W. New Neoverrucosane Diterpenoids Produced by the Marine Gliding Bacterium Saprospira Grandis. *J. Nat. Prod.* **2003**, *66* (6), 818–822. https://doi.org/10.1021/np0205351.

(172) Yamamoto, T.; Izumi, N.; Ui, H.; Sueki, A.; Masuma, R.; Nonaka, K.; Hirose, T.; Sunazuka, T.; Nagai, T.; Yamada, H.; Ōmura, S.; Shiomi, K. Wickerols A and B: Novel Anti-Influenza Virus Diterpenes Produced by Trichoderma Atroviride FKI-3849. *Tetrahedron* **2012**, *68* (45), 9267–9271. https://doi.org/10.1016/j.tet.2012.08.066.

(173) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. **2014**, *4* (5), 468–481. https://doi.org/10.1002/wcms.1183.

(174) Lanzoni, O.; Sabaneyeva, E.; Modeo, L.; Castelli, M.; Lebedeva, N.; Verni, F.; Schrallhammer, M.; Potekhin, A.; Petroni, G. Diversity and Environmental Distribution of the Cosmopolitan Endosymbiont " Candidatus Megaira." *Sci. Rep.* **2019**, *9* (1), 1179. https://doi.org/10.1038/s41598-018-37629-w.

(175) Zhu, G.; Hou, C.; Yuan, W.; Wang, Z.; Zhang, J.; Jiang, L.; Karthik, L.; Li, B.; Ren, B.; Lv, K.; Lu, W.; Cong, Z.; Dai, H.; Hsiang, T.; Zhang, L.; Liu, X. Molecular Networking Assisted Discovery and Biosynthesis Elucidation of the Antimicrobial Spiroketals Epicospirocins. *Chem. Commun.* **2020**. https://doi.org/10.1039/D0CC03990J.

(176) Cheng, X.; Liang, X.; Zheng, Z.-H.; Zhang, X.-X.; Lu, X.-H.; Yao, F.-H.; Qi, S.-H. Penicimeroterpenoids A–C, Meroterpenoids with Rearrangement Skeletons from the Marine-Derived Fungus Penicillium Sp. SCSIO 41512. *Org. Lett.* **2020**. https://doi.org/10.1021/acs.orglett.0c02160.

(177) Kwon, Y.; Shin, J.; Nam, K.; An, J. S.; Yang, S.-H.; Hong, S.-H.; Bae, M.; Moon, K.; Cho, Y.; Woo, J.; Park, K.; Kim, K.; Shin, J.; Kim, B.-Y.; Kim, Y.; Oh, D.-C. Rhizolutin, a Novel 7/10/6-Tricyclic Dilactone, Dissociates Misfolded Protein Aggregates and Reduces Apoptosis/Inflammation Associated with Alzheimer's Disease. *Angew. Chem. Int. Ed.* **2020**. https://doi.org/10.1002/anie.202009294.

(178) Xu, Z. F.; Bo, S. T.; Wang, M. J.; Shi, J.; Jiao, R. H.; Sun, Y.; Xu, Q.; Tan, R.; Ge, H. M. Discovery and Biosynthesis of Bosamycin from Streptomyces Sp. 120454. *Chem. Sci.* **2020**. https://doi.org/10.1039/D0SC03469J.

(179) Luyen, N. D.; Huong, L. M.; Thi Hong Ha, T.; Cuong, L. H.; Thi Hai Yen, D.; Nhiem, N. X.; Tai, B. H.; Gardes, A.; Kopprio, G.; Van Kiem, P. Aspermicrones A-C, Novel Dibenzospiroketals from the Seaweed-Derived Endophytic Fungus Aspergillus Micronesiensis. *J. Antibiot. (Tokyo)* **2019**, *72* (11), 843–847. https://doi.org/10.1038/s41429-019-0214-8.

(180) Kosemura, S. Meroterpenoids from Penicillium Citreo-Viride B. IFO 4692 and 6200 Hybrid. *Tetrahedron* **2003**, *59* (27), 5055–5072. https://doi.org/10.1016/S0040-4020(03)00739-7.

(181) Endo, A. Monacolin K, a New Hypocholesterolemic Agent That Specifically Inhibits 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase. *J. Antibiot. (Tokyo)* **1980**, *33* (3), 334–336. https://doi.org/10.7164/antibiotics.33.334.

(182) Ji, G.; Beavis, R.; Novick, R. P. Bacterial Interference Caused by Autoinducing Peptide Variants. *Science* **1997**, *276* (5321), 2027–2030. https://doi.org/10.1126/science.276.5321.2027.

(183) Wu, Y.; Liao, H.; Liu, L.-Y.; Sun, F.; Chen, H.-F.; Jiao, W.-H.; Zhu, H.-R.; Yang, F.; Huang, G.; Zeng, D.-Q.; Zhou, M.; Wang, S.-P.; Lin, H.-W. Phakefustatins A–C: Kynurenine-Bearing Cycloheptapeptides as RXRα Modulators from the Marine Sponge Phakellia Fusca. *Org. Lett.* **2020**. https://doi.org/10.1021/acs.orglett.0c01586.

(184) Naman, C. B.; Rattan, R.; Nikoulina, S. E.; Lee, J.; Miller, B. W.; Moss, N. A.; Armstrong, L.; Boudreau, P. D.; Debonsi, H. M.; Valeriote, F. A.; Dorrestein, P. C.; Gerwick, W. H. Integrating Molecular Networking and Biological Assays To Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an American Samoan Marine Cyanobacterium. *J. Nat. Prod.* **2017**, *80* (3), 625–633. https://doi.org/10.1021/acs.jnatprod.6b00907.

(185) Brinkmann, C. M.; Marker, A.; Kurtböke, D. İ. An Overview on Marine Sponge-Symbiotic Bacteria as Unexhausted Sources for Natural Product Discovery. *Diversity* **2017**, *9* (4), 40. https://doi.org/10.3390/d9040040.

(186) Han, M.; Liu, F.; Zhang, F.; Li, Z.; Lin, H. Bacterial and Archaeal Symbionts in the South China Sea Sponge Phakellia Fusca: Community Structure, Relative Abundance, and Ammonia-Oxidizing Populations. *Mar. Biotechnol. N. Y. N* **2012**, *14* (6), 701–713. https://doi.org/10.1007/s10126-012-9436-5.

(187) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. *J. Cheminformatics* **2020**, *12* (1), 20. https://doi.org/10.1186/s13321-020-00424-9.
(188) Chen, Y.; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inform.* **2020**. https://doi.org/10.1002/minf.202000171.
(189) Dias, D. A.; Urban, S.; Roessner, U. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2012**, *2* (2), 303–336. https://doi.org/10.3390/metabo2020303.
(190) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.
(191) 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.
(192) Zaid, H.; Raiyn, J.; Nasser, A.; Saad, B.; Rayan, A. Physicochemical Properties of Natural Based Products versus Synthetic Chemicals. *Open Nutraceuticals J.* **2010**, *3* (1).
(193) Yu, M. J. Natural Product-Like Virtual Libraries: Recursive Atom-Based Enumeration. *J. Chem. Inf. Model.* **2011**, *51* (3), 541–557. https://doi.org/10.1021/ci1002087.
(194) 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.
(195) Pereira, F. Machine Learning Methods to Predict the Terrestrial and Marine Origin of Natural Products. *Mol. Inform.* **2021**, *n/a* (n/a). https://doi.org/10.1002/minf.202060034.
(196) 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. Cheminformatics* **2016**, *8* (1), 61. https://doi.org/10.1186/s13321-016-0174-y.
(197) Kim, H.; Wang, M.; Leber, C.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P.; Gerwick, W.; Cottrell, G. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. **2020**. https://doi.org/10.26434/chemrxiv.12885494.v1.
(198) Meunier, L.; Tocquin, P.; Cornet, L.; Sirjacobs, D.; Leclère, V.; Pupin, M.; Jacques, P.; Baurain, D. Palantir: A Springboard for the Analysis of Secondary Metabolite Gene Clusters in Large-Scale Genome Mining Projects. *Bioinformatics* **2020**, *36* (15), 4345–4347. https://doi.org/10.1093/bioinformatics/btaa517.
(199) *Horizontal Gene Transfer: Breaking Borders Between Living Kingdoms*; Villa, T. G., Viñas, M., Eds.; Springer International Publishing: Cham, 2019. https://doi.org/10.1007/978-3-030-21862-1.
(200) Hardoim, P. R.; van Overbeek, L. S.; Berg, G.; Pirttilä, A. M.; Compant, S.; Campisano, A.; Döring, M.; Sessitsch, A. The Hidden World within Plants: Ecological and Evolutionary Considerations for Defining Functioning of Microbial Endophytes. *Microbiol. Mol. Biol. Rev. MMBR* **2015**, *79* (3), 293–320. https://doi.org/10.1128/MMBR.00050-14.
(201) Strobel, G.; Daisy, B.; Castillo, U.; Harper, J. Natural Products from Endophytic Microorganisms. *J. Nat. Prod.* **2004**, *67* (2), 257–268. https://doi.org/10.1021/np030397v.
(202) Ye, K.; Ai, H.-L.; Liu, J.-K. Identification and Bioactivities of Secondary Metabolites Derived from Endophytic Fungi Isolated from Ethnomedicinal Plants of Tujia in Hubei Province: A Review. *Nat. Prod. Bioprospecting* **2021**, *11* (2), 185–205. https://doi.org/10.1007/s13659-020-00295-5.
(203) Howat, S.; Park, B.; Oh, I. S.; Jin, Y.-W.; Lee, E.-K.; Loake, G. J. Paclitaxel: Biosynthesis, Production and Future Prospects. *New Biotechnol.* **2014**, *31* (3), 242–245. https://doi.org/10.1016/j.nbt.2014.02.010.
(204) Shankar Naik, B. Developments in Taxol Production through Endophytic Fungal Biotechnology: A Review. *Orient. Pharm. Exp. Med.* **2019**, *19* (1), 1–13. https://doi.org/10.1007/s13596-018-0352-8.
(205) Kusari, S.; Lamshöft, M.; Kusari, P.; Gottfried, S.; Zühlke, S.; Louven, K.; Hentschel, U.; Kayser, O.; Spiteller, M. Endophytes Are Hidden Producers of Maytansine in Putterlickia Roots. *J. Nat. Prod.* **2014**, *77* (12), 2577–2584. https://doi.org/10.1021/np500219a.

(206) Heim, W. G.; Sykes, K. A.; Hildreth, S. B.; Sun, J.; Lu, R.-H.; Jelesko, J. G. Cloning and Characterization of a Nicotiana Tabacum Methylputrescine Oxidase Transcript. *Phytochemistry* **2007**, *68* (4), 454–463. https://doi.org/10.1016/j.phytochem.2006.11.003.
(207) Hooven, H. W. van den; Lagerwerf, F. M.; Heerma, W.; Haverkamp, J.; Piard, J.-C.; Hilbers, C. W.; Siezen, R. J.; Kuipers, O. P.; Rollema, H. S. The Structure of the Lantibiotic Lacticin 481 Produced by Lactococcus Lactis: Location of the Thioether Bridges. *FEBS Lett.* **1996**, *391* (3), 317–322. https://doi.org/10.1016/0014-5793(96)00771-5.
(208) Cao, P.-R.; Zheng, Y.-L.; Zhao, Y.-Q.; Wang, X.-B.; Zhang, H.; Zhang, M.-H.; Yang, T.; Gu, Y.-C.; Yang, M.-H.; Kong, L.-Y. Beetleane A and Epicoane A: Two Carbon Skeletons Produced by Epicoccum Nigrum. *Org. Lett.* **2021**. https://doi.org/10.1021/acs.orglett.1c00731.
(209) Zhang, J.; Yuan, M.-F.; Li, S.-T.; Sang, C.-C.; Chen, M.-F.; Ao, Y.-L.; Li, Z.-W.; Xie, J.; Ye, W.-C.; Zhang, X.-Q. Hunzeylanines A–E, Five Bisindole Alkaloids Tethered with a Methylene Group from the Roots of Hunteria Zeylanica. *J. Org. Chem.* **2020**, *85* (16), 10884–10890. https://doi.org/10.1021/acs.joc.0c01448.
(210) Lou, H.; Yi, P.; Hu, Z.; Li, Y.; Zeng, Y.; Gu, W.; Huang, L.; Yuan, C.; Hao, X. Polycyclic Polyprenylated Acylphloroglucinols with Acetylcholinesterase Inhibitory Activities from Hypericum Perforatum. *Fitoterapia* **2020**, *143*, 104550. https://doi.org/10.1016/j.fitote.2020.104550.
(211) Li, S.-G.; Wang, Y.-T.; Zhang, Q.; Wang, K.-B.; Xue, J.-J.; Li, D.-H.; Jing, Y.-K.; Lin, B.; Hua, H.-M. Pegaharmols A–B, Axially Chiral β-Carboline-Quinazoline Dimers from the Roots of Peganum Harmala. *Org. Lett.* **2020**, *22* (19), 7522–7525. https://doi.org/10.1021/acs.orglett.0c02709.
(212) Zhang, J.; Shi, L.-Y.; Yin, X.; Xu, F.-C.; Zhang, Q.-Y.; Tu, P.-F.; Liang, H. Discovery of Novel Potential Plant Growth Regulators from Corydalis Mucronifera. *Fitoterapia* **2020**, *147*, 104776. https://doi.org/10.1016/j.fitote.2020.104776.
(213) Wu, J.; Zhao, S.-M.; Shi, B.-B.; Bao, M.-F.; Schinnerl, J.; Cai, X.-H. Cage-Monoterpenoid Quinoline Alkaloids with Neurite Growth Promoting Effects from the Fruits of Melodinus Yunnanensis. *Org. Lett.* **2020**, *22* (19), 7676–7680. https://doi.org/10.1021/acs.orglett.0c02871.
(214) Tanaka, N.; Niwa, K.; Kajihara, S.; Tsuji, D.; Itoh, K.; Mamadalieva, N. Z.; Kashiwada, Y. C28 Terpenoids from Lamiaceous Plant Perovskia Scrophulariifolia: Their Structures and Anti-Neuroinflammatory Activity. *Org. Lett.* **2020**, *22* (19), 7667–7670. https://doi.org/10.1021/acs.orglett.0c02855.
(215) Fan, Y.-Y.; Gan, L.-S.; Chen, S.-X.; Gong, Q.; Zhang, H.-Y.; Yue, J.-M. Horienoids A and B, Two Heterocoupled Sesquiterpenoid Dimers from Hedyosmum Orientale. *J. Org. Chem.* **2021**. https://doi.org/10.1021/acs.joc.1c00307.
(216) Kim, M. C.; Winter, J. M.; Asolkar, R. N.; Boonlarppradab, C.; Cullum, R.; Fenical, W. Marinoterpins A–C: Rare Linear Merosesterterpenoids from Marine-Derived Actinomycete Bacteria of the Family Streptomycetaceae. *J. Org. Chem.* **2021**. https://doi.org/10.1021/acs.joc.1c00262.
(217) Wu, P.-L.; Hsu, Y.-L.; Jao, C.-W. Indole Alkaloids from Cephalanceropsis Gracilis. *J. Nat. Prod.* **2006**, *69* (10), 1467–1470. https://doi.org/10.1021/np0603951.
(218) Mason, J. J.; Bergman, J.; Janosik, T. Synthetic Studies of Cephalandole Alkaloids and the Revised Structure of Cephalandole A. *J. Nat. Prod.* **2008**, *71* (8), 1447–1450. https://doi.org/10.1021/np800334j.
(219) Ishikura, M.; Yamada, K. Simple Indole Alkaloids and Those with a Nonrearranged Monoterpenoid Unit. *Nat. Prod. Rep.* **2009**, *26* (6), 803–852. https://doi.org/10.1039/B820693G.
(220) Zhao, J.; Zhou, L.-L.; Li, X.; Xiao, H.-B.; Hou, F.-F.; Cheng, Y.-X. Bioactive Compounds from the Aerial Parts of Brachystemma Calycinum and Structural Revision of an Octacyclopeptide. *J. Nat. Prod.* **2011**, *74* (6), 1392–1400. https://doi.org/10.1021/np200048u.
(221) Yeshak, M. Y.; Burman, R.; Asres, K.; Göransson, U. Cyclotides from an Extreme Habitat: Characterization of Cyclic Peptides from Viola Abyssinica of the Ethiopian Highlands. *J. Nat. Prod.* **2011**, *74* (4), 727–731. https://doi.org/10.1021/np100790f.

(222) Srivastava, S.; Dashora, K.; Ameta, K. L.; Singh, N. P.; El-Enshasy, H. A.; Pagano, M. C.; Hesham, A. E.-L.; Sharma, G. D.; Sharma, M.; Bhargava, A. Cysteine-Rich Antimicrobial Peptides from Plants: The Future of Antimicrobial Therapy. *Phytother. Res.* **2021**, *35* (1), 256–277. https://doi.org/10.1002/ptr.6823.

(223) Santos-Silva, C. A. dos; Zupin, L.; Oliveira-Lima, M.; Vilela, L. M. B.; Bezerra-Neto, J. P.; Ferreira-Neto, J. R.; Ferreira, J. D. C.; Oliveira-Silva, R. L. de; Pires, C. de J.; Aburjaile, F. F.; Oliveira, M. F. de; Kido, E. A.; Crovella, S.; Benko-Iseppon, A. M. Plant Antimicrobial Peptides: State of the Art, In Silico Prediction and Perspectives in the Omics Era. *Bioinforma. Biol. Insights* **2020**, *14*, 1177932220952739. https://doi.org/10.1177/1177932220952739.

(224) Tursch, B.; Braekman, J. C.; Daloze, D.; Herin, M.; Karlsson, R. Chemical Studies of Marine Invertebrates. X. Lobophytolide, a New Cembranolide Diterpene from the Soft Coral Lobophytum Cristagalli (Coelenterata, Octocorallia, Alcyonacea). *Tetrahedron Lett.* **1974**, *15* (43), 3769–3772. https://doi.org/10.1016/S0040-4039(01)92004-0.

(225) Blunt, J. W.; Copp, B. R.; Munro, M. H. G.; Northcote, P. T.; Prinsep, M. R. Marine Natural Products. *Nat. Prod. Rep.* **2010**, *27* (2), 165–237. https://doi.org/10.1039/B906091J.

(226) Ovenden, S. P. B.; Capon, R. J. Echinosulfonic Acids A−C and Echinosulfone A: Novel Bromoindole Sulfonic Acids and a Sulfone from a Southern Australian Marine Sponge, Echinodictyum. *J. Nat. Prod.* **1999**, *62* (9), 1246–1249. https://doi.org/10.1021/np9901027.

(227) Dhinakaran, D. I.; Prasad, D. R. D.; Gohila, R.; Lipton, P. Screening of Marine Sponge-Associated Bacteria from Echinodictyum Gorgonoides and Its Bioactivity. *Afr. J. Biotechnol.* **2012**, *11* (88), 15469–15476. https://doi.org/10.4314/ajb.v11i88.

(228) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18* (8), 1093–1110. https://doi.org/10.1016/j.neunet.2005.07.009.

(229) Sakula, A. Paul Langerhans (1847-1888): A Centenary Tribute. *J. R. Soc. Med.* **1988**, *81* (7), 414–415.

(230) Kingsberg, S. A.; Clayton, A. H.; Portman, D.; Williams, L. A.; Krop, J.; Jordan, R.; Lucas, J.; Simon, J. A. Bremelanotide for the Treatment of Hypoactive Sexual Desire Disorder. *Obstet. Gynecol.* **2019**, *134* (5), 899–908. https://doi.org/10.1097/AOG.0000000000003500.

(231) Al Shaer, D.; Al Musaimi, O.; Albericio, F.; de la Torre, B. G. 2019 FDA TIDES (Peptides and Oligonucleotides) Harvest. *Pharmaceuticals* **2020**, *13* (3), 40. https://doi.org/10.3390/ph13030040.

(232) de la Torre, B. G.; Albericio, F. The Pharmaceutical Industry in 2019. An Analysis of FDA Drug Approvals from the Perspective of Molecules. *Molecules* **2020**, *25* (3), 745. https://doi.org/10.3390/molecules25030745.

(233) Lam, K. S. Affinity Selection and Sequencing. *Nat. Chem. Biol.* **2019**, *15* (4), 320–321. https://doi.org/10.1038/s41589-019-0253-2.

(234) Fjell, C. D.; Hiss, J. A.; Hancock, R. E. W.; Schneider, G. Designing Antimicrobial Peptides: Form Follows Function. *Nat. Rev. Drug Discov.* **2012**, *11* (1), 37–51. https://doi.org/10.1038/nrd3591.

(235) Mansbach, R. A.; Travers, T.; McMahon, B. H.; Fair, J. M.; Gnanakaran, S. Snails In Silico: A Review of Computational Studies on the Conopeptides. *Mar. Drugs* **2019**, *17* (3), 145. https://doi.org/10.3390/md17030145.

(236) Torres, M. D. T.; Sothiselvam, S.; Lu, T. K.; Fuente-Nunez, C. Peptide Design Principles for Antimicrobial Applications. *J. Mol. Biol.* **2019**. https://doi.org/10.1016/j.jmb.2018.12.015.

(237) Mulligan, V. K. The Emerging Role of Computational Design in Peptide Macrocycle Drug Discovery. *Expert Opin. Drug Discov.* **2020**, *15* (7), 833–852. https://doi.org/10.1080/17460441.2020.1751117.

(238) Lin, X.; Li, X.; Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **2020**, *25* (6), 1375. https://doi.org/10.3390/molecules25061375.

(239) Nikiforovich, G. V. Computational Molecular Modeling in Peptide Drug Design. *Int. J. Pept. Protein Res.* **1994**, *44* (6), 513–531. https://doi.org/10.1111/j.1399-3011.1994.tb01140.x.

(240) Lee, A. C.-L.; Harris, J. L.; Khanna, K. K.; Hong, J.-H. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *Int. J. Mol. Sci.* **2019**, *20* (10), 2383. https://doi.org/10.3390/ijms20102383.

(241) Han, Y.; Král, P. Computational Design of ACE2-Based Peptide Inhibitors of SARS-CoV-2. *ACS Nano* **2020**, *14* (4), 5143–5147. https://doi.org/10.1021/acsnano.0c02857.
(242) Sevy, A. M.; Gilchuk, I. M.; Brown, B. P.; Bozhanova, N. G.; Nargi, R.; Jensen, M.; Meiler, J.; Crowe, J. E. Computationally Designed Cyclic Peptides Derived from an Antibody Loop Increase Breadth of Binding for Influenza Variants. *Structure* **2020**, *28* (10), 1114-1123.e4. https://doi.org/10.1016/j.str.2020.04.005.
(243) Korendovych, I. V.; DeGrado, W. F. De Novo Protein Design, a Retrospective. *Q. Rev. Biophys.* **2020**, *53*. https://doi.org/10.1017/S0033583519000131.
(244) Cao, L.; Goreshnik, I.; Coventry, B.; Case, J. B.; Miller, L.; Kozodoy, L.; Chen, R. E.; Carter, L.; Walls, A. C.; Park, Y.-J.; Strauch, E.-M.; Stewart, L.; Diamond, M. S.; Veesler, D.; Baker, D. De Novo Design of Picomolar SARS-CoV-2 Miniprotein Inhibitors. *Science* **2020**, *370* (6515), 426–431. https://doi.org/10.1126/science.abd9909.
(245) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P.-S.; Huber, T.; Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.; Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D.-A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, R. Y.-R.; Watkins, A.; Zimmerman, L.; Bonneau, R. Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks. *Nat. Methods* **2020**, *17* (7), 665–680. https://doi.org/10.1038/s41592-020-0848-2.
(246) Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Mao, B.; Foight, G. W.; Lee, M. Y.; Gagnon, L. A.; Carter, L.; Sankaran, B.; Ovchinnikov, S.; Marcos, E.; Huang, P.-S.; Vaughan, J. C.; Stoddard, B. L.; Baker, D. De Novo Design of a Fluorescence-Activating β-Barrel. *Nature* **2018**, *561* (7724), 485–491. https://doi.org/10.1038/s41586-018-0509-0.
(247) Kang, S.-M.; Moon, H.; Han, S.-W.; Kim, D.-H.; Kim, B. M.; Lee, B.-J. Structure-Based De Novo Design of Mycobacterium Tuberculosis VapC-Activating Stapled Peptides. *ACS Chem. Biol.* **2020**, *15* (9), 2493–2498. https://doi.org/10.1021/acschembio.0c00492.
(248) Zhang, S.-Q.; Huang, H.; Yang, J.; Kratochvil, H. T.; Lolicato, M.; Liu, Y.; Shu, X.; Liu, L.; DeGrado, W. F. Designed Peptides That Assemble into Cross-α Amyloid-like Structures. *Nat. Chem. Biol.* **2018**, *14* (9), 870–875. https://doi.org/10.1038/s41589-018-0105-5.
(249) Holland, J. H. Genetic Algorithms. *Sci. Am.* **1992**, *267* (1), 66–73.
(250) Teixidó, M.; Belda, I.; Roselló, X.; González, S.; Fabre, M.; Llorá, X.; Bacardit, J.; Garrell, J. M.; Vilaró, S.; Albericio, F.; Giralt, E. Development of a Genetic Algorithm to Design and Identify Peptides That Can Cross the Blood-Brain Barrier. *QSAR Comb. Sci.* **2003**, *22* (7), 745–753. https://doi.org/10.1002/qsar.200320004.
(251) Beltran, J. A.; Brizuela, C. A. Design of Selective Cationic Antibacterial Peptides: A Multiobjective Genetic Algorithm Approach. In *2016 IEEE Congress on Evolutionary Computation (CEC)*; IEEE: Vancouver, BC, Canada, 2016; pp 484–491. https://doi.org/10.1109/CEC.2016.7743833.
(252) Porto, W. F.; Irazazabal, L.; Alves, E. S. F.; Ribeiro, S. M.; Matos, C. O.; Pires, Á. S.; Fensterseifer, I. C. M.; Miranda, V. J.; Haney, E. F.; Humblot, V.; Torres, M. D. T.; Hancock, R. E. W.; Liao, L. M.; Ladram, A.; Lu, T. K.; Fuente-Nunez, C.; Franco, O. L. In Silico Optimization of a Guava Antimicrobial Peptide Enables Combinatorial Exploration for Peptide Design. *Nat. Commun.* **2018**, *9*. https://doi.org/10.1038/s41467-018-03746-3.

(253) Knapp, B.; Giczi, V.; Ribarics, R.; Schreiner, W. PeptX: Using Genetic Algorithms to Optimize Peptides for MHC Binding. *BMC Bioinformatics* **2011**, *12*, 241. https://doi.org/10.1186/1471-2105-12-241.
(254) King, M. D.; Long, T.; Andersen, T.; McDougal, O. M. Genetic Algorithm Managed Peptide Mutant Screening: Optimizing Peptide Ligands for Targeted Receptor Binding. *J. Chem. Inf. Model.* **2016**, *56* (12), 2378–2387. https://doi.org/10.1021/acs.jcim.6b00095.
(255) Fjell, C. D.; Jenssen, H.; Cheung, W. A.; Hancock, R. E. W.; Cherkasov, A. Optimization of Antibacterial Peptides by Genetic Algorithms and Cheminformatics. *Chem. Biol. Drug Des.* **2011**, *77* (1), 48–56. https://doi.org/10.1111/j.1747-0285.2010.01044.x.
(256) Yoshida, M.; Hinkley, T.; Tsuda, S.; Abul-Haija, Y. M.; McBurney, R. T.; Kulikov, V.; Mathieson, J. S.; Galiñanes Reyes, S.; Castro, M. D.; Cronin, L. Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem* **2018**, *4* (3), 533–543. https://doi.org/10.1016/j.chempr.2018.01.005.
(257) Neuhaus, C. S.; Gabernet, G.; Steuer, C.; Root, K.; Hiss, J. A.; Zenobi, R.; Schneider, G. Simulated Molecular Evolution for Anticancer Peptide Design. *Angew. Chem. Int. Ed.* **2019**, *58* (6), 1674–1678. https://doi.org/10.1002/anie.201811215.
(258) Basith, S.; Manavalan, B.; Shin, T. H.; Lee, G. Machine Intelligence in Peptide Therapeutics: A next-Generation Tool for Rapid Disease Screening. *Med. Res. Rev.* **2020**, *40* (4), 1276–1314. https://doi.org/10.1002/med.21658.
(259) Aranha, M. P.; Spooner, C.; Demerdash, O.; Czejdo, B.; Smith, J. C.; Mitchell, J. C. Prediction of Peptide Binding to MHC Using Machine Learning with Sequence and Structure-Based Feature Sets. *Biochim. Biophys. Acta BBA - Gen. Subj.* **2020**, *1864* (4), 129535. https://doi.org/10.1016/j.bbagen.2020.129535.
(260) Zhang, Y. P.; Zou, Q. PPTPP: A Novel Therapeutic Peptide Prediction Method Using Physicochemical Property Encoding and Adaptive Feature Representation Learning. *Bioinformatics* **2020**, *36* (13), 3982–3987. https://doi.org/10.1093/bioinformatics/btaa275.
(261) Lee, E. Y.; Fulan, B. M.; Wong, G. C. L.; Ferguson, A. L. Mapping Membrane Activity in Undiscovered Peptide Sequence Space Using Machine Learning. *Proc. Natl. Acad. Sci.* **2016**, *113* (48), 13588–13593. https://doi.org/10.1073/pnas.1609893113.
(262) Spänig, S.; Heider, D. Encodings and Models for Antimicrobial Peptide Classification for Multi-Resistant Pathogens. *BioData Min.* **2019**, *12* (1), 7. https://doi.org/10.1186/s13040-019-0196-x.
(263) 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.
(264) Cherkasov, A.; Hilpert, K.; Jenssen, H.; Fjell, C. D.; Waldbrook, M.; Mullaly, S. C.; Volkmer, R.; Hancock, R. E. W. Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chem. Biol.* **2009**, *4* (1), 65–74. https://doi.org/10.1021/cb800240j.
(265) Timmons, P. B.; Hewage, C. M. HAPPENN Is a Novel Tool for Hemolytic Activity Prediction for Therapeutic Peptides Which Employs Neural Networks. *Sci. Rep.* **2020**, *10* (1), 10869. https://doi.org/10.1038/s41598-020-67701-3.
(266) Gautam, A.; Chaudhary, K.; Singh, S.; Joshi, A.; Anand, P.; Tuknait, A.; Mathur, D.; Varshney, G. C.; Raghava, G. P. S. Hemolytik: A Database of Experimentally Determined Hemolytic and Non-Hemolytic Peptides. *Nucleic Acids Res.* **2014**, *42* (Database issue), D444-449. https://doi.org/10.1093/nar/gkt1008.
(267) Müller, A. T.; Hiss, J. A.; Schneider, G. Recurrent Neural Network Model for Constructive Peptide Design. *J. Chem. Inf. Model.* **2018**, *58* (2), 472–479. https://doi.org/10.1021/acs.jcim.7b00414.
(268) Tucs, A.; Tran, D. P.; Yumoto, A.; Ito, Y.; Uzawa, T.; Tsuda, K. Generating Ampicillin-Level Antimicrobial Peptides with Activity-Aware Generative Adversarial Networks. *ACS Omega* **2020**, *5* (36), 22847–22851. https://doi.org/10.1021/acsomega.0c02088.

(269) Nguyen, L. T.; Haney, E. F.; Vogel, H. J. The Expanding Scope of Antimicrobial Peptide Structures and Their Modes of Action. *Trends Biotechnol.* **2011**, *29* (9), 464–472. https://doi.org/10.1016/j.tibtech.2011.05.001.
(270) Stach, M.; Siriwardena, T. N.; Kohler, T.; Delden, C.; Darbre, T.; Reymond, J. L. Combining Topology and Sequence Design for the Discovery of Potent Antimicrobial Peptide Dendrimers against Multidrug-Resistant Pseudomonas Aeruginosa. *Angew Chem Int Ed Engl* **2014**, *53* (47), 12827–12831. https://doi.org/10.1002/anie.201409270.
(271) Gan, B.-H.; Siriwardena, T. N.; Javor, S.; Darbre, T.; Reymond, J.-L. Fluorescence Imaging of Bacterial Killing by Antimicrobial Peptide Dendrimer G3KL. *ACS Infect. Dis.* **2019**, *5* (12), 2164–2173. https://doi.org/10.1021/acsinfecdis.9b00299.
(272) Jeddou, F. B.; Falconnet, L.; Luscher, A.; Siriwardena, T.; Reymond, J.-L.; Delden, C. van; Köhler, T. Adaptive and Mutational Responses to Peptide Dendrimer Antimicrobials in Pseudomonas Aeruginosa. *Antimicrob. Agents Chemother.* **2020**, *64* (4). https://doi.org/10.1128/AAC.02040-19.
(273) Abdel-Sayed, P.; Kaeppeli, A.; Siriwardena, T.; Darbre, T.; Perron, K.; Jafari, P.; Reymond, J. L.; Pioletti, D. P.; Applegate, L. A. Anti-Microbial Dendrimers against Multidrug-Resistant P. Aeruginosa Enhance the Angiogenic Effect of Biological Burn-Wound Bandages. *Sci Rep* **2016**, *6*, 1–10. https://doi.org/10.1038/srep22020.
(274) Han, X.; Liu, Y.; Ma, Y.; Zhang, M.; He, Z.; Siriwardena, T. N.; Xu, H.; Bai, Y.; Zhang, X.; Reymond, J.-L.; Qiao, M. Peptide Dendrimers G3KL and TNS18 Inhibit Pseudomonas Aeruginosa Biofilms. *Appl. Microbiol. Biotechnol.* **2019**, *103* (14), 5821–5830. https://doi.org/10.1007/s00253-019-09801-3.
(275) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.
(276) Singh, S.; Chaudhary, K.; Dhanda, S. K.; Bhalla, S.; Usmani, S. S.; Gautam, A.; Tuknait, A.; Agrawal, P.; Mathur, D.; Raghava, G. P. S. SATPdb: A Database of Structurally Annotated Therapeutic Peptides. *Nucleic Acids Res.* **2016**, *44* (D1), D1119–D1126. https://doi.org/10.1093/nar/gkv1114.
(277) Pirtskhalava, M.; Gabrielian, A.; Cruz, P.; Griggs, H. L.; Squires, R. B.; Hurt, D. E.; Grigolava, M.; Chubinidze, M.; Gogoladze, G.; Vishnepolsky, B.; Alekseev, V.; Rosenthal, A.; Tartakovsky, M. DBAASP v.2: An Enhanced Database of Structure and Antimicrobial/Cytotoxic Activity of Natural and Synthetic Peptides. *Nucleic Acids Res.* **2016**, *44* (Database issue), D1104–D1112. https://doi.org/10.1093/nar/gkv1174.
(278) Kang, X.; Dong, F.; Shi, C.; Liu, S.; Sun, J.; Chen, J.; Li, H.; Xu, H.; Lao, X.; Zheng, H. DRAMP 2.0, an Updated Data Repository of Antimicrobial Peptides. *Sci. Data* **2019**, *6*. https://doi.org/10.1038/s41597-019-0154-y.
(279) Qureshi, A.; Thakur, N.; Tandon, H.; Kumar, M. AVPdb: A Database of Experimentally Validated Antiviral Peptides Targeting Medically Important Viruses. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1147-1153. https://doi.org/10.1093/nar/gkt1191.
(280) Choo, K. H.; Tan, T. W.; Ranganathan, S. SPdb – a Signal Peptide Database. *BMC Bioinformatics* **2005**, *6* (1), 249. https://doi.org/10.1186/1471-2105-6-249.
(281) Kim, Y.; Bark, S.; Hook, V.; Bandeira, N. NeuroPedia: Neuropeptide Database and Spectral Library. *Bioinforma. Oxf. Engl.* **2011**, *27* (19), 2772–2773. https://doi.org/10.1093/bioinformatics/btr445.
(282) Novković, M.; Simunić, J.; Bojović, V.; Tossi, A.; Juretić, D. DADP: The Database of Anuran Defense Peptides. *Bioinforma. Oxf. Engl.* **2012**, *28* (10), 1406–1407. https://doi.org/10.1093/bioinformatics/bts141.
(283) Wynendaele, E.; Bronselaer, A.; Nielandt, J.; D'Hondt, M.; Stalmans, S.; Bracke, N.; Verbeke, F.; Van De Wiele, C.; De Tré, G.; De Spiegeleer, B. Quorumpeps Database: Chemical Space, Microbial Origin and Functionality of Quorum Sensing Peptides. *Nucleic Acids Res.* **2013**, *41* (Database issue), D655–D659. https://doi.org/10.1093/nar/gks1137.
(284) Ettayapuram Ramaprasad, A. S.; Singh, S.; Gajendra P S, R.; Venkatesan, S. AntiAngioPred: A Server for Prediction of Anti-Angiogenic Peptides. *PloS One* **2015**, *10* (9), e0136990. https://doi.org/10.1371/journal.pone.0136990.
(285) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432* (7019), 823–823.

(286) Deng, Z.-L.; Du, C.-X.; Li, X.; Hu, B.; Kuang, Z.-K.; Wang, R.; Feng, S.-Y.; Zhang, H.-Y.; Kong, D.-X. Exploring the Biologically Relevant Chemical Space for Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53* (11), 2820–2828. https://doi.org/10.1021/ci400432a.
(287) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**. https://doi.org/10.1016/j.drudis.2019.02.013.
(288) Medina-Franco, J. L.; Aguayo-Ortiz, R. Progress in the Visualization and Mining of Chemical and Target Spaces. *Mol Inf* **2013**, *32* (11–12), 942–953. https://doi.org/10.1002/minf.201300041.
(289) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opin Drug Discov* **2015**, *10* (9), 959–973. https://doi.org/10.1517/17460441.2015.1060216.
(290) Lin, A.; Horvath, D.; Afonina, V.; Marcou, G.; Reymond, J.-L.; Varnek, A. Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds. *ChemMedChem* **2018**, *13* (6), 540–554. https://doi.org/doi:10.1002/cmdc.201700561.
(291) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem Sci* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/c8sc00148k.
(292) van Hilten, N.; Chevillard, F.; Kolb, P. Virtual Compound Libraries in Computer-Assisted Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 644–651. https://doi.org/10.1021/acs.jcim.8b00737.
(293) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J Cheminf* **2017**, *9* (1), 48. https://doi.org/10.1186/s13321-017-0235-x.
(294) Schneider, G. Automating Drug Discovery. *Nat. Rev. Drug Discov.* **2018**, *17* (2), 97–113. https://doi.org/10.1038/nrd.2017.232.
(295) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23* (6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.
(296) Henninot, A.; Collins, J. C.; Nuss, J. M. The Current State of Peptide Drug Discovery: Back to the Future? *J. Med. Chem.* **2018**, *61* (4), 1382–1414. https://doi.org/10.1021/acs.jmedchem.7b00318.
(297) Morrison, C. Constrained Peptides' Time to Shine? *Nat Rev Drug Discov* **2018**, *17*, 531–533. https://doi.org/10.1038/nrd.2018.125.
(298) Reymond, J. L.; Ruddigkeit, L.; Blum, L. C.; Van Deursen, R. The Enumeration of Chemical Space. *WIREs Comput Mol Sci* **2012**, *2* (5), 713–733.
(299) Krause, T.; Röckendorf, N.; El-Sourani, N.; Ramaker, K.; Henkel, M.; Hauke, S.; Borschbach, M.; Frey, A. Breeding Cell Penetrating Peptides: Optimization of Cellular Uptake by a Function-Driven Evolutionary Process. *Bioconjug. Chem.* **2018**. https://doi.org/10.1021/acs.bioconjchem.8b00583.
(300) Haney, E. F.; Brito-Sánchez, Y.; Trimble, M. J.; Mansour, S. C.; Cherkasov, A.; Hancock, R. E. W. Computer-Aided Discovery of Peptides That Specifically Attack Bacterial Biofilms. *Sci. Rep.* **2018**, *8* (1), 1871. https://doi.org/10.1038/s41598-018-19669-4.
(301) Grisoni, F.; Neuhaus, C. S.; Hishinuma, M.; Gabernet, G.; Hiss, J. A.; Kotera, M.; Schneider, G. De Novo Design of Anticancer Peptides by Ensemble Artificial Neural Networks. *J. Mol. Model.* **2019**, *25* (5), 112. https://doi.org/10.1007/s00894-019-4007-6.
(302) Wu, Q.; Ke, H.; Li, D.; Wang, Q.; Fang, J.; Zhou, J. Recent Progress in Machine Learning-Based Prediction of Peptide Activity for Drug Discovery. *Curr Top Med Chem* **2019**, *19* (1), 4–16. https://doi.org/10.2174/1568026619666190122151634.
(303) Gabernet, G.; Gautschi, D.; Müller, A. T.; Neuhaus, C. S.; Armbrecht, L.; Dittrich, P. S.; Hiss, J. A.; Schneider, G. In Silico Design and Optimization of Selective Membranolytic Anticancer Peptides. *Sci Rep* **2019**, *9* (1), 1–11. https://doi.org/10.1038/s41598-019-47568-9.

(304) Kalafatovic, D.; Mauša, G.; Todorovski, T.; Giralt, E. Algorithm-Supported, Mass and Sequence Diversity-Oriented Random Peptide Library Design. *J Cheminf* **2019**, *11* (1), 25. https://doi.org/10.1186/s13321-019-0347-6.
(305) Sauer, W. H.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J Chem Inf Comput Sci* **2003**, *43* (3), 987–1003. https://doi.org/10.1021/ci025599w.
(306) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J Med Chem* **2010**, *53* (10), 3862–3886. https://doi.org/10.1021/jm900818s.
(307) Morstein, J.; Awale, M.; Reymond, J.-L.; Trauner, D. Mapping the Azolog Space Enables the Optical Control of New Biological Targets. *ACS Cent. Sci.* **2019**, *5* (4), 607–618. https://doi.org/10.1021/acscentsci.8b00881.
(308) Selsted, M. E.; Novotny, M. J.; Morris, W. L.; Tang, Y. Q.; Smith, W.; Cullor, J. S. Indolicidin, a Novel Bactericidal Tridecapeptide Amide from Neutrophils. *J. Biol. Chem.* **1992**, *267* (7), 4292–4295.
(309) Dubos, R. J.; Hotchkiss, R. D. The Production of Bactericidal Substances by Aerobic Sporulating Bacilli. *J. Exp. Med.* **1941**, *73* (5), 629–640. https://doi.org/10.1084/jem.73.5.629.
(310) Jin, A.-H.; Muttenthaler, M.; Dutertre, S.; Himaya, S. W. A.; Kaas, Q.; Craik, D. J.; Lewis, R. J.; Alewood, P. F. Conotoxins: Chemistry and Biology. *Chem. Rev.* **2019**. https://doi.org/10.1021/acs.chemrev.9b00207.
(311) Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Dokl. Akad. Nauk SSSR* **1965**, *163*, 845–848.
(312) Jácome, D.; Tapia, F.; Lascano, J. E.; Fuertes, W. Contextual Analysis of Comments in B2C Facebook Fan Pages Based on the Levenshtein Algorithm. In *Information Technology and Systems*; Rocha, Á., Ferrás, C., Paredes, M., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing, 2019; pp 528–538.
(313) Crooks, G. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14* (6), 1188–1190. https://doi.org/10.1101/gr.849004.
(314) Naylor, M. R.; Bockus, A. T.; Blanco, M.-J.; Lokey, R. S. Cyclic Peptide Natural Products Chart the Frontier of Oral Bioavailability in the Pursuit of Undruggable Targets. *Curr Opin Chem Biol* **2017**, *38*, 141–147. https://doi.org/10.1016/j.cbpa.2017.04.012.
(315) Zorzi, A.; Deyle, K.; Heinis, C. Cyclic Peptide Therapeutics: Past, Present and Future. *Curr Opin Chem Biol* **2017**, *38*, 24–29. http://dx.doi.org/10.1016/j.cbpa.2017.02.006.
(316) Yang, X.; Yousef, A. E. Antimicrobial Peptides Produced by Brevibacillus Spp.: Structure, Classification and Bioactivity: A Mini Review. *World J Microbiol Biotechnol* **2018**, *34* (4), 57. https://doi.org/10.1007/s11274-018-2437-4.
(317) Goodman, M.; Chorev, M. On the Concept of Linear Modified Retro-Peptide Structures. *Acc. Chem. Res.* **1979**, *12* (1), 1–7. https://doi.org/10.1021/ar50133a001.
(318) Arranz-Gibert, P.; Ciudad, S.; Seco, J.; García, J.; Giralt, E.; Teixidó, M. Immunosilencing Peptides by Stereochemical Inversion and Sequence Reversal: Retro-D-Peptides. *Sci Rep* **2018**, *8* (1), 6446. https://doi.org/10.1038/s41598-018-24517-6.
(319) McGivern, J. G. Ziconotide: A Review of Its Pharmacology and Use in the Treatment of Pain. *Neuropsychiatr Treat* **2007**, *3* (1), 69–85.
(320) Reymond, J.-L.; Darbre, T. Peptide and Glycopeptide Dendrimer Apple Trees as Enzyme Models and for Biomedical Applications. *Org. Biomol. Chem.* **2012**, *10* (8), 1483–1492. https://doi.org/10.1039/c2ob06938e.
(321) Maillard, N.; Clouet, A.; Darbre, T.; Reymond, J. L. Combinatorial Libraries of Peptide Dendrimers: Design, Synthesis, on-Bead High-Throughput Screening, Bead Decoding and Characterization. *Nat Protoc* **2009**, *4* (2), 132–142. https://doi.org/10.1038/nprot.2008.241.
(322) Höfle, G.; Bedorf, N.; Steinmetz, H.; Schomburg, D.; Gerth, K.; Reichenbach, H. Epothilone A and B—Novel 16-Membered Macrolides with Cytotoxic Activity: Isolation, Crystal Structure, and Conformation in Solution. *Angew Chem Int Ed Engl* **1996**, *35* (13–14), 1567–1569. https://doi.org/10.1002/anie.199615671.

(323) Chen, C.; Posocco, P.; Liu, X.; Cheng, Q.; Laurini, E.; Zhou, J.; Liu, C.; Wang, Y.; Tang, J.; Col, V. D.; Yu, T.; Giorgio, S.; Fermeglia, M.; Qu, F.; Liang, Z.; Rossi, J. J.; Liu, M.; Rocchi, P.; Pricl, S.; Peng, L. Mastering Dendrimer Self-Assembly for Efficient SiRNA Delivery: From Conceptual Design to In Vivo Efficient Gene Silencing. *Small* **2016**, *12* (27), 3667–3676. https://doi.org/10.1002/smll.201503866.

(324) Heitz, M.; Javor, S.; Darbre, T.; Reymond, J.-L. Stereoselective PH Responsive Peptide Dendrimers for SiRNA Transfection. *Bioconjug. Chem.* **2019**, *30* (8), 2165–2182. https://doi.org/10.1021/acs.bioconjchem.9b00403.

(325) ChemAxon - Software Solutions and Services for Chemistry & Biology https://chemaxon.com/ (accessed 2018 -09 -25).

(326) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(327) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

(328) Bian, Y.; Xie, X.-Q. Generative Chemistry: Drug Discovery with Deep Learning Generative Models. *J. Mol. Model.* **2021**, *27* (3), 71. https://doi.org/10.1007/s00894-021-04674-8.

(329) Cardoso, M. H.; Orozco, R. Q.; Rezende, S. B.; Rodrigues, G.; Oshiro, K. G. N.; Cândido, E. S.; Franco, O. L. Computer-Aided Design of Antimicrobial Peptides: Are We Generating Effective Drug Candidates? *Front. Microbiol.* **2020**, *10*. https://doi.org/10.3389/fmicb.2019.03097.

(330) Lee, E. Y.; Lee, M. W.; Fulan, B. M.; Ferguson, A. L.; Wong, G. C. L. What Can Machine Learning Do for Antimicrobial Peptides, and What Can Antimicrobial Peptides Do for Machine Learning? *Interface Focus* **2017**, *7* (6), 20160153. https://doi.org/10.1098/rsfs.2016.0153.

(331) Piotto, S. P.; Sessa, L.; Concilio, S.; Iannelli, P. YADAMP: Yet Another Database of Antimicrobial Peptides. *Int. J. Antimicrob. Agents* **2012**, *39* (4), 346–351. https://doi.org/10.1016/j.ijantimicag.2011.12.003.

(332) Gogoladze, G.; Grigolava, M.; Vishnepolsky, B.; Chubinidze, M.; Duroux, P.; Lefranc, M.-P.; Pirtskhalava, M. DBAASP: Database of Antimicrobial Activity and Structure of Peptides. *FEMS Microbiol. Lett.* **2014**, *357* (1), 63–68. https://doi.org/10.1111/1574-6968.12489.

(333) Lee, H.-T.; Lee, C.-C.; Yang, J.-R.; Lai, J. Z. C.; Chang, K. Y. A Large-Scale Structural Classification of Antimicrobial Peptides. *BioMed Res. Int.* **2015**, *2015*, e475062. https://doi.org/10.1155/2015/475062.

(334) Wang, G.; Li, X.; Wang, Z. APD3: The Antimicrobial Peptide Database as a Tool for Research and Education. *Nucleic Acids Res.* **2016**, *44* (D1), D1087–D1093. https://doi.org/10.1093/nar/gkv1278.

(335) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. CAMPR3: A Database on Sequences, Structures and Signatures of Antimicrobial Peptides. *Nucleic Acids Res.* **2016**, *44* (Database issue), D1094–D1097. https://doi.org/10.1093/nar/gkv1051.

(336) Ye, G.; Wu, H.; Huang, J.; Wang, W.; Ge, K.; Li, G.; Zhong, J.; Huang, Q. LAMP2: A Major Update of the Database Linking Antimicrobial Peptides. *Database* **2020**, *2020* (baaa061). https://doi.org/10.1093/database/baaa061.

(337) Santajit, S.; Indrawattana, N. Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *BioMed Res. Int.* **2016**, *2016*. https://doi.org/10.1155/2016/2475067.

(338) Mulani, M. S.; Kamble, E. E.; Kumkar, S. N.; Tawre, M. S.; Pardesi, K. R. Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. *Front. Microbiol.* **2019**, *10*. https://doi.org/10.3389/fmicb.2019.00539.

(339) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; Cherkasov, A.; Seleem, M. N.; Pinilla, C.; de la Fuente-Nunez, C.; Lazaridis, T.; Dai, T.; Houghten, R. A.; Hancock, R. E. W.; Tegos, G. P. The Value of Antimicrobial Peptides in the Age of Resistance. *Lancet Infect. Dis.* **2020**, *20* (9), e216–e230. https://doi.org/10.1016/S1473-3099(20)30327-3.

(340) Browne, K.; Chakraborty, S.; Chen, R.; Willcox, M. D.; Black, D. S.; Walsh, W. R.; Kumar, N. A New Era of Antibiotics: The Clinical Potential of Antimicrobial Peptides. *Int. J. Mol. Sci.* **2020**, *21* (19), 7047. https://doi.org/10.3390/ijms21197047.

(341) Baeriswyl, S.; Gan, B.-H.; Siriwardena, T. N.; Visini, R.; Robadey, M.; Javor, S.; Stocker, A.; Darbre, T.; Reymond, J.-L. X-Ray Crystal Structures of Short Antimicrobial Peptides as Pseudomonas Aeruginosa Lectin B Complexes. *ACS Chem. Biol.* **2019**, *14* (4), 758–766. https://doi.org/10.1021/acschembio.9b00047.

(342) Greco, I.; Molchanova, N.; Holmedal, E.; Jenssen, H.; Hummel, B. D.; Watts, J. L.; Håkansson, J.; Hansen, P. R.; Svenson, J. Correlation between Hemolytic Activity, Cytotoxicity and Systemic in Vivo Toxicity of Synthetic Antimicrobial Peptides. *Sci. Rep.* **2020**, *10* (1), 13206. https://doi.org/10.1038/s41598-020-69995-9.

(343) Schneider, P.; Müller, A. T.; Gabernet, G.; Button, A. L.; Posselt, G.; Wessler, S.; Hiss, J. A.; Schneider, G. Hybrid Network Model for "Deep Learning" of Chemical Data: Application to Antimicrobial Peptides. *Mol. Inform.* **2017**, *36* (1–2), 1600011. https://doi.org/10.1002/minf.201600011.

(344) Meher, P. K.; Sahu, T. K.; Saini, V.; Rao, A. R. Predicting Antimicrobial Peptides with Improved Accuracy by Incorporating the Compositional, Physico-Chemical and Structural Features into Chou's General PseAAC. *Sci. Rep.* **2017**, *7* (1), 42362. https://doi.org/10.1038/srep42362.

(345) Liu, S.; Bao, J.; Lao, X.; Zheng, H. Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides. *Sci. Rep.* **2018**, *8* (1), 11189. https://doi.org/10.1038/s41598-018-29566-5.

(346) Veltri, D.; Kamath, U.; Shehu, A. Deep Learning Improves Antimicrobial Peptide Recognition. *Bioinformatics* **2018**, *34* (16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179.

(347) Su, X.; Xu, J.; Yin, Y.; Quan, X.; Zhang, H. Antimicrobial Peptide Identification Using Multi-Scale Convolutional Network. *BMC Bioinformatics* **2019**, *20* (1), 730. https://doi.org/10.1186/s12859-019-3327-y.

(348) Vishnepolsky, B.; Zaalishvili, G.; Karapetian, M.; Nasrashvili, T.; Kuljanishvili, N.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M.; Grigolava, M.; Pirtskhalava, M. De Novo Design and In Vitro Testing of Antimicrobial Peptides against Gram-Negative Bacteria. *Pharmaceuticals* **2019**, *12* (2), 82. https://doi.org/10.3390/ph12020082.

(349) Yan, J.; Bhadra, P.; Li, A.; Sethiya, P.; Qin, L.; Tai, H. K.; Wong, K. H.; Siu, S. W. I. Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther. - Nucleic Acids* **2020**, *20*, 882–894. https://doi.org/10.1016/j.omtn.2020.05.006.

(350) Nagarajan, D.; Nagarajan, T.; Roy, N.; Kulkarni, O.; Ravichandran, S.; Mishra, M.; Chakravortty, D.; Chandra, N. Computational Antimicrobial Peptide Design and Evaluation against Multidrug-Resistant Clinical Isolates of Bacteria. *J. Biol. Chem.* **2018**, *293* (10), 3492–3509. https://doi.org/10.1074/jbc.M117.805499.

(351) Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; dos Santos, C.; Chen, P.-Y.; Yang, Y. Y.; Tan, J. P. K.; Hedrick, J.; Crain, J.; Mojsilovic, A. Accelerated Antimicrobial Discovery via Deep Generative Models and Molecular Dynamics Simulations. *Nat. Biomed. Eng.* **2021**, 1–11. https://doi.org/10.1038/s41551-021-00689-x.

(352) Maccari, G.; Luca, M. D.; Nifosí, R.; Cardarelli, F.; Signore, G.; Boccardi, C.; Bifone, A. Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization. *PLOS Comput. Biol.* **2013**, *9* (9), e1003212. https://doi.org/10.1371/journal.pcbi.1003212.

(353) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The Helical Hydrophobic Moment: A Measure of the Amphiphilicity of a Helix. *Nature* **1982**, *299* (5881), 371–374. https://doi.org/10.1038/299371a0.

(354) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(355) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. **2019**, 12.

(356) Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv14061078 Cs Stat* **2014**.
(357) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
(358) Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. In *International Conference on Machine Learning*; 2013; pp 1139–1147.
(359) Haapala, A. *Ztane/Python-Levenshtein*; 2020.
(360) Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-Sequence-Based Prediction of Protein Secondary Structures and Solvent Accessibility by Deep Whole-Sequence Learning. *J. Comput. Chem.* **2018**, *39* (26), 2210–2216. https://doi.org/10.1002/jcc.25534.
(361) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS Force Field. *Eur. Biophys. J.* **2005**, *34* (4), 273–284. https://doi.org/10.1007/s00249-004-0448-6.
(362) Chiu, S. W.; Clark, M.; Balaji, V.; Subramaniam, S.; Scott, H. L.; Jakobsson, E. Incorporation of Surface Tension into Molecular Dynamics Simulation of an Interface: A Fluid Phase Lipid Bilayer Membrane. *Biophys. J.* **1995**, *69* (4), 1230–1245.
(363) Comeau, C.; Ries, B.; Stadelmann, T.; Tremblay, J.; Poulet, S.; Fröhlich, U.; Côté, J.; Boudreault, P.-L.; Derbali, R. M.; Sarret, P.; Grandbois, M.; Leclair, G.; Riniker, S.; Marsault, É. Modulation of the Passive Permeability of Semipeptidic Macrocycles: N- and C-Methylations Fine-Tune Conformation and Properties. *J. Med. Chem.* **2021**, *64* (9), 5365–5383. https://doi.org/10.1021/acs.jmedchem.0c02036.
(364) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. **2021**. https://doi.org/10.26434/chemrxiv.14233418.v1.
(365) Siriwardena, T. N.; Gan, B.-H.; Köhler, T.; van Delden, C.; Javor, S.; Reymond, J.-L. Stereorandomization as a Method to Probe Peptide Bioactivity. *ACS Cent. Sci.* **2021**, *7* (1), 126–134. https://doi.org/10.1021/acscentsci.0c01135.
(366) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577* (7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7.
(367) Kornfeld, S.; Li, E.; Tabas, I. The Synthesis of Complex-Type Oligosaccharides. II. Characterization of the Processing Intermediates in the Synthesis of the Complex Oligosaccharide Units of the Vesicular Stomatitis Virus G Protein. *J. Biol. Chem.* **1978**, *253* (21), 7771–7778. https://doi.org/10.1016/S0021-9258(17)34436-8.
(368) Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegenthart, J. F.; Lütteke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **2015**, *25* (12), 1323–1324. https://doi.org/10.1093/glycob/cwv091.
(369) 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.

Appendix A – Supplementary Tables and Figures for Chapter Two

Note A.1: Table 21

Table 21. Examples of SMILES strings that give error or not meaningful analogs using PubChem and/or ChEMBL search.

name	SMILES	Error description
Exenatide CID 45588096	CC[C@H](C)[C@@H](C(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CC1=CNC2=CC=CC=C21)C(=O)N[C@@H](CC(C)C)C(=O)N[C@@H](CCCCN)C(=O)N[C@@H](CC(=O)N)C(=O)NCC(=O)NCC(=O)N3CCC[C@H]3C(=O)N[C@@H](CO)C(=O)N[C@@H](CO)C(=O)NCC(=O)N[C@@H](C)C(=O)N4CCC[C@H]4C(=O)N5CCC[C@H]5C(=O)N6CCC[C@H]6C(=O)N[C@@H](CO)C(=O)N)NC(=O)[C@H](CC7=CC=CC=C7)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCNC(=N)N)NC(=O)[C@H](C(C)C)NC(=O)[C@H](C)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](CCSC)NC(=O)[C@H](CCC(=O)N)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CO)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CC(=O)O)NC(=O)[C@H](CO)NC(=O)[C@H]([C@@H](C)O)NC(=O)[C@H](CC8=CC=CC=C8)NC(=O)[C@H]([C@@H](C)O)NC(=O)CNC(=O)[C@H](CCC(=O)O)NC(=O)CNC(=O)[C@H](CC9=CNC=N9)N	PuChem search ignores size ans shape, as the first results are all small peptides.
Mipomersen sodium CID 118984460	CC1=CN(C(=O)NC1=O)[C@H]2C[C@@H]([C@H](O2)COP(=S)([O-])O[C@H]3C[C@@H](O[C@@H]3COP(=S)([O-])O[C@H]4C[C@@H](O[C@@H]4COP(=S)([O-])O[C@H]5C[C@@H](O[C@@H]5COP(=S)([O-])O[C@H]6C[C@@H](O[C@@H]6COP(=S)([O-])O[C@H]7C[C@@H](O[C@@H]7COP(=S)([O-])O[C@H]8C[C@@H](O[C@@H]8COP(=S)([O-])O[C@H]9C[C@@H](O[C@@H]9COP(=S)([O-])O[C@H]1C[C@@H](O[C@@H]1COP(=S)([O-])O[C@@H]1[C@H](O[C@H]([C@@H]1OCCOC)N1C=C(C(=NC1=O)N)C)COP(=S)([O-])O[C@@H]1[C@H](O[C@H]([C@@H]1OCCOC)N1C=C(C(=O)NC1=O)C)COP(=S)([O-])O[C@@H]1[C@H](O[C@H]([C@@H]1OCCOC)N1C=C(C(=NC1=O)N)C)COP(=S)([O-])O[C@@H]1[C@H](O[C@H]([C@@H]1OCCOC)N1C=C(C(=NC1=O)N)C)COP(=S)([O-])O[C@@H]1[C@H](O[C@H]([C@@H]1OCCOC)N1C=NC2=C1N=C(NC2=O)N)CO)N1C=NC2=C1N=CN=C2N)N1C=NC2=C1N=C(NC2=O)N)N1C=C(C(=O)NC1=O)C)N1C=C(C(=NC1=O)N)C)N1C=C(C(=O)NC1=O)C)N1C=NC2=C1N=C(NC2=O)N)N1C=C(C(=NC1=O)N)C)N1C=C(C(=O)NC1=O)C)OP(=S)([O-])OC[C@@H]1[C@H](C[C@@H](O1)N1C=C(C(=NC1=O)N)C)OP(=S)([O-])OC[C@@H]1[C@H]([C@H]([C@@H](O1)N1C=NC2=C1N=C(NC2=O)N)OCCOC)OP(=S)([O-])OC[C@@H]1[C@H]([C@H]([C@@H](O1)N1C=C(C(=NC1=O)N)C)OCCOC)OP(=S)([O-	PubChem search ignores size ans shape, as the first results are all small oligonucleotides.

	])OC[C@@H]1[C@H]([C@H]([C@@H](O1)N1C=NC2=C1N=CN=C2N)OCCOC)OP(=O)(OC[C@@H]1[C@H]([C@H]([C@@H](O1)N1C=C(C(=NC1=O)N)C)OCCOC)OP(=S)([O-])OC[C@@H]1[C@H]([C@H]([C@@H](O1)N1C=C(C(=NC1=O)N)C)OCCOC)O)[S-].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+].[Na+]	
G3KL (KL)8(KKL)4 (KKL)2KKL	CC(C)C[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)C(=O)O	PubChem search fails. ChEMBL search ignores shape, as the frst results are linear peptides.
CID 89996683	C(COCCN(CCOCCN)CC(=O)NCCOCCN(CCOCCNC(=O)CN(CCOCCN)CCOCCN)CC(=O)NCCOCCN(CCOCCNC(=O)CN(CCOCCNC(=O)CN(CCOCCN)CCOCCN)CCOCCNC(=O)CN(CCOCCN)CCOCCN)CCOCCNC(=O)CN(CCOCCN)CCOCCN)CC(=O)NCCOCCN(CCOCCNC(=O)CN(CCOCCNC(=O)CN(CCOCCNC(=O)CN(CCOCCN)CCOCCN)CCOCCNC(=O)CN(CCOCCN)CCOCCN)CCOCCNC(=O)CN(CCOCCN)CCOCCN)CCOCCNC(=O)CN(CCOCCN)CCOCCN)CC(=O)O)N	PubChem search fails. ChEMBL search ignores dendrimer shape.
Maitotoxin 1	C[C@H](CC[C@@H]([C@@H]([C@H](C)C[C@H](C(=C)/C(=C/CO)/C)O)O)OS(=O)(=O)[O-])[C@H]([C@@H](C)[C@H]1[C@@H]([C@@H]([C@H]2[C@H](O1)[C@@H](C[C@]3([C@H](O2)C[C@H]4[C@H](O3)C[C@]5([C@H](O4)[C@H]([C@H]6[C@H](O5)C[C@H]([C@H](O6)[C@@H]([C@H](C[C@H]7[C@@H]([C@@H]([C@H]8[C@H](O7)C[C@H]9[C@H](O8)C[C@H]1[C@H](O9)[C@H]([C@@H]2[C@@H](O1)[C@@H]([C@H]([C@@H](O2)[C@H]1[C@@H]([C@H]([C@H]2[C@@H](O1)C[C@H]([C@@H](O2)[C@@H](C[C@H](C[C@H]1[C@@H]([C@H]([C@H]2[C@@H](O1)C[C@H]([C@@H](O2)[C@H]1[C@@H](C[C@]2([C@H](O1)[C@@H]([C@]1([C@H](O2)C[C@]2([C@H](O1)CC[C@]1([C@H](O2)C[C@]2([C@H](O1)C[C@H]1[C@H](O2)CC[C@H](O1)[C@]1([C@@H](C[C@H]2[C@](O1)(C[C@H]1[C@](O2)(CC[C@]2([C@H](O1)C[C@H]1[C@](O2)(C[C@H]2[C@H](O1)C/C=C\[C@H]1[C@H](O2)C[C@H]2[C@](O1)(C[C@]1([C@H](O2)C[C@H]2[C@](O1)(CC[C@H](O2)[C@H]([C@@H](C[C@@H](C)[C@@H](C)CC=C)O)O)C)C)C)C)C)C)C)C)C)O)C)O)O)O)O)O)O)O)O)O)O)O)O)O)OS(=O)(=O)[O-])O)O)O)O)C)C)O)O)O)O.[Na+].[Na+]	PubChem search doesn't find similar compounds.

GalAXG3	O=C([C@H](CCCCNC([C@@H](NC([C@@H](NC([C@H](CCCCNC([C@H]1N(C([C@@H](NC(CSC[C@H](NC([C@H]([C@@H](C)CC)NC([C@H](CCCCNC([C@@H]2CCCN2C([C@@H](NC(C3=CC=C(O[C@H]4[C@H](O)[C@@H](O)[C@@H](O)[C@@H](CO)O4)C=C3)=O)CCCC[NH3+])=O)=O)NC([C@@H]5CCCN5C([C@@H](NC(C6=CC=C(O[C@H]7[C@H](O)[C@@H](O)[C@@H](O)[C@@H](CO)O7)C=C6)=O)CCCC[NH3+])=O)=O)=O)=O)C(N)=O)=O)CCCC[NH3+])=O)CCC1)=O)NC([C@H]8N(C([C@@H](NC(CSC[C@H](NC([C@H]([C@@H](C)CC)NC([C@H](CCCCNC([C@@H]9CCCN9C([C@@H](NC(C%10=CC=C(O[C@@H]%11O[C@H](CO)[C@H](O)[C@H](O)[C@H]%11O)C=C%10)=O)CCCC[NH3+])=O)=O)NC([C@@H]%12CCCN%12C([C@@H](NC(C%13=CC=C(O[C@@H]%14O[C@H](CO)[C@H](O)[C@H](O)[C@H]%14O)C=C%13)=O)CCCC[NH3+])=O)=O)=O)=O)C(N)=O)=O)CCCC[NH3+])=O)CCC8)=O)=O)CCCC[NH3+])=O)[C@@H](C)CC)=O)NC([C@@H](NC([C@@H](NC([C@H](CCCCNC([C@H]%15N(C([C@@H](NC(CSC[C@H](NC([C@H]([C@@H](C)CC)NC([C@H](CCCCNC([C@@H]%16CCCN%16C([C@@H](NC(C%17=CC=C(O[C@@H]%18O[C@H](CO)[C@H](O)[C@H](O)[C@H]%18O)C=C%17)=O)CCCC[NH3+])=O)=O)NC([C@@H]%19CCCN%19C([C@@H](NC(C%20=CC=C(O[C@@H]%21O[C@H](CO)[C@H](O)[C@H](O)[C@H]%21O)C=C%20)=O)CCCC[NH3+])=O)=O)=O)=O)C(N)=O)=O)CCCC[NH3+])=O)CCC%15)=O)NC([C@H]%22N(C([C@@H](NC(CSC[C@H](NC([C@H]([C@@H](C)CC)NC([C@H](CCCCNC([C@@H]%23CCCN%23C([C@@H](NC(C%24=CC=C(O[C@H]%25[C@H](O)[C@@H](O)[C@@H](O)[C@@H](CO)O%25)C=C%24)=O)CCCC[NH3+])=O)=O)NC([C@@H]%26CCCN%26C([C@@H](NC(C%27=CC=C(O[C@H]%28[C@H](O)[C@@H](O)[C@@H](O)[C@@H](CO)O%28)C=C%27)=O)CCCC[NH3+])=O)=O)=O)=O)C(N)=O)=O)CCCC[NH3+])=O)CCC%22)=O)=O)CCCC[NH3+])=O)[C@@H](C)CC)=O)N[C@@H](CC%29=CN=CN%29)C(N[C@@H]([C@H](CC)C)C(N)=O)=O	PubChem search fails.
T7	CC(C)C[C@H](NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CCCCN)NC(=O)[C@H](CCCCNC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)NC(=O)[C@H](CC(C)C)NC(=O)[C@@H](N)CCCCN)C(=O)O	Both PubChem and ChEMBL searches ignore dendrimer shape, as the first results are linear peptides.

Vivagel	OS(=O)(=O)C1=CC=C2C(OCC(=O)NCCCC[C@H](NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)C(=O)NCCCC[C@H](NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)C(=O)NCCCC[C@H](NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)C(=O)NCCCC[C@H](NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)C(=O)NCCCC[C@H](NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)[C@H](CCCCNC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)NC(=O)COC3=CC(=CC4=CC(=CC=C34)S(O)(=O)=O)S(O)(=O)=O)C(=O)NC(C3=CC=CC=C3)C3=CC=CC=C3)=CC(=CC2=C1)S(O)(=O)=O	Both PubChem and ChEMBL searches fail.

Appendix B – Supplementary Tables and Figures for Chapter Three

Note B.1: Table 22

Table 22. Linear random peptide sequences used to generate the mutated and scrambled peptide datasets (available at https://github.com/reymond-group/map4) for the extended fingerprint benchmark.

Length	Random sequences
10 residues	KAQIDLSPNP
	TITVVSLMNQ
	SSQHVQRENY
	PNWKLRPNYH
	QHVEYEQHDL
	LQKLNDTWCT
	HRIWAWVNMN
	VNWVWDHFSR
	IDIIIRTHES
	LFCCVAYSFD
20 residues	NLCADQWYFSNAFW
	VIHGPWLIQGISHAI
	DHVRWIFWHAAICCD
	RTVSFSYCRQDCQPF
	KGHYERMCEPVRKKN
	NNYYPKYREPQGKII
	VGPYVKHPAEACNQQ
	QETTHQIPFMTAFAH
	RYAWTRHFAIFVVDV
	FGDWRAGYGGSPENL
30 residues	IRCWIWVECLYINHRVHEYW
	HQDAICAETWIVKWATLNSW
	HWYEKCCHGAQSWWTVDQWL
	RWPWKPFHVRFFFQRWPLLH
	FPGFIAKCTQWGAGLEVGGH
	SPYDQEMDQQFDRACCWHFK
	VSNWRLLGPRPYMSNTETQR
	NAPSYKTFEANITTRAARNS
	VPMMTDGIMTVYVGGEPWSR
	HAETEGYKSSPTKLGCDPLY

Note B.2: Figures 42-48

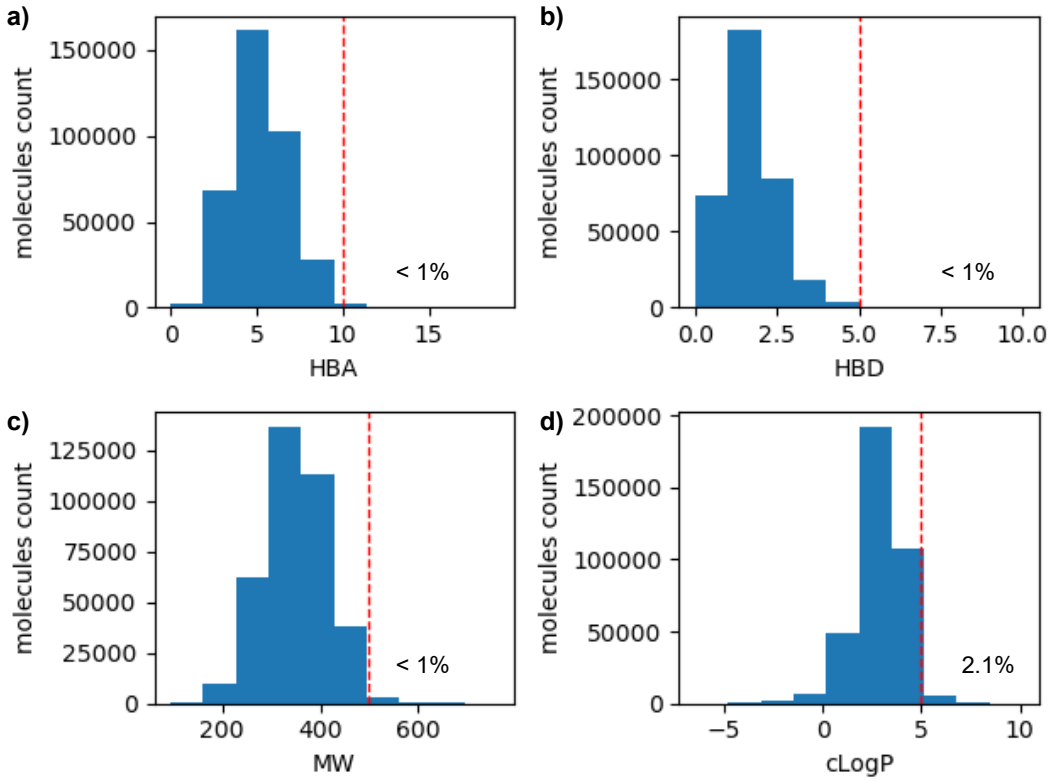


Figure 38. Hydrogen bond acceptor (HBA, a) and donor (HBD, b) count, Molecular Weight (MW, c), and calculated logarithm octanol-water partition coefficient (cLogP, d) of the actives/decoys used in the original version of the Riniker fingerprint benchmark. The percentage of molecules that violate each rule is reported in the corresponded panel.

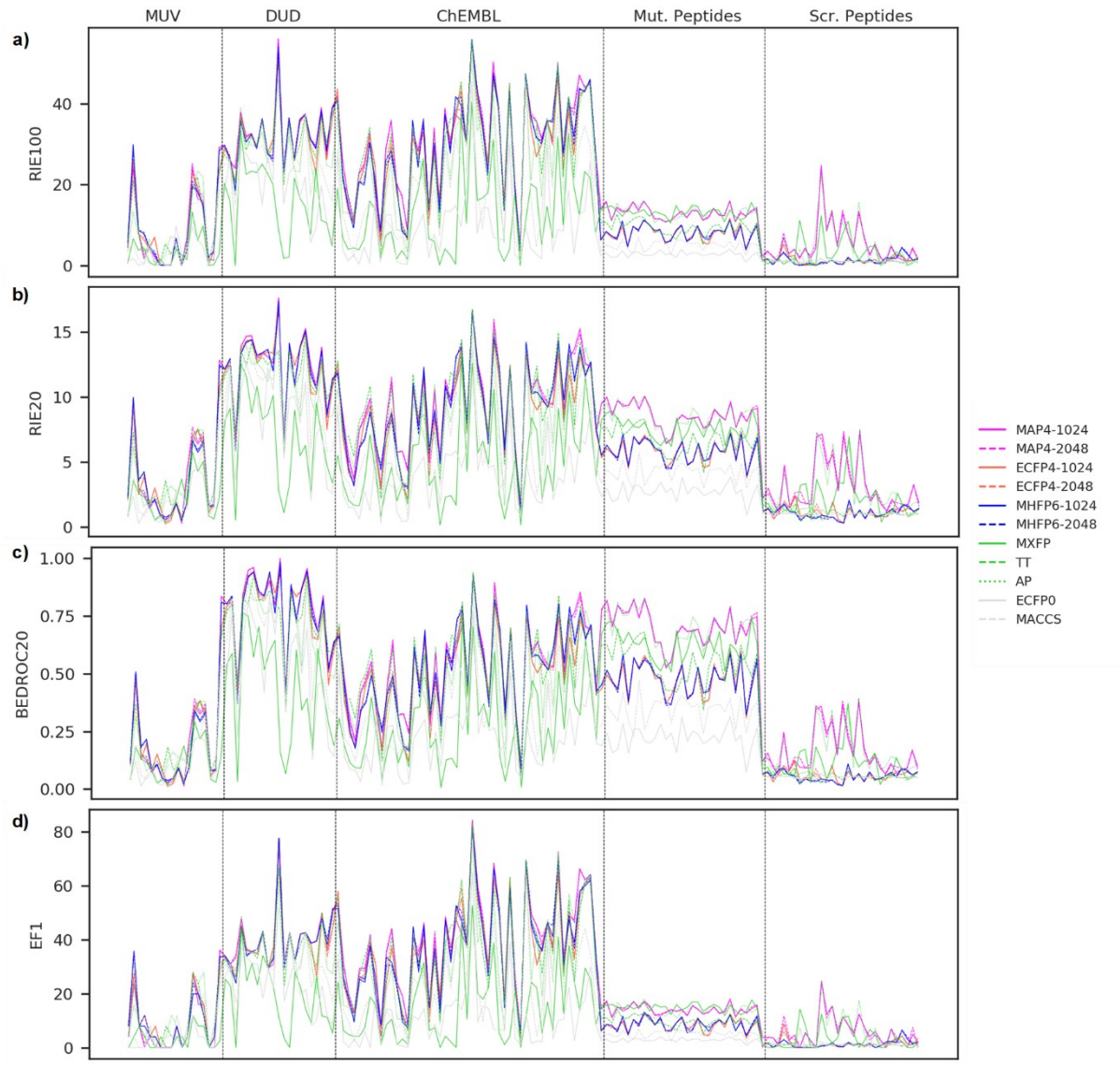


Figure 39. RIE100 (a), RIE20 (b), BEDROC20 (c), and EF1 (d) of MAP4 (magenta), ECFP4 (orange), MHFP6 (blue), MXFP (solid green line), TT (dashed green line), AP (dotted green line), MACCS (solid gray line), and ECFP0 (dashed gray line) across all small molecules and peptide targets.

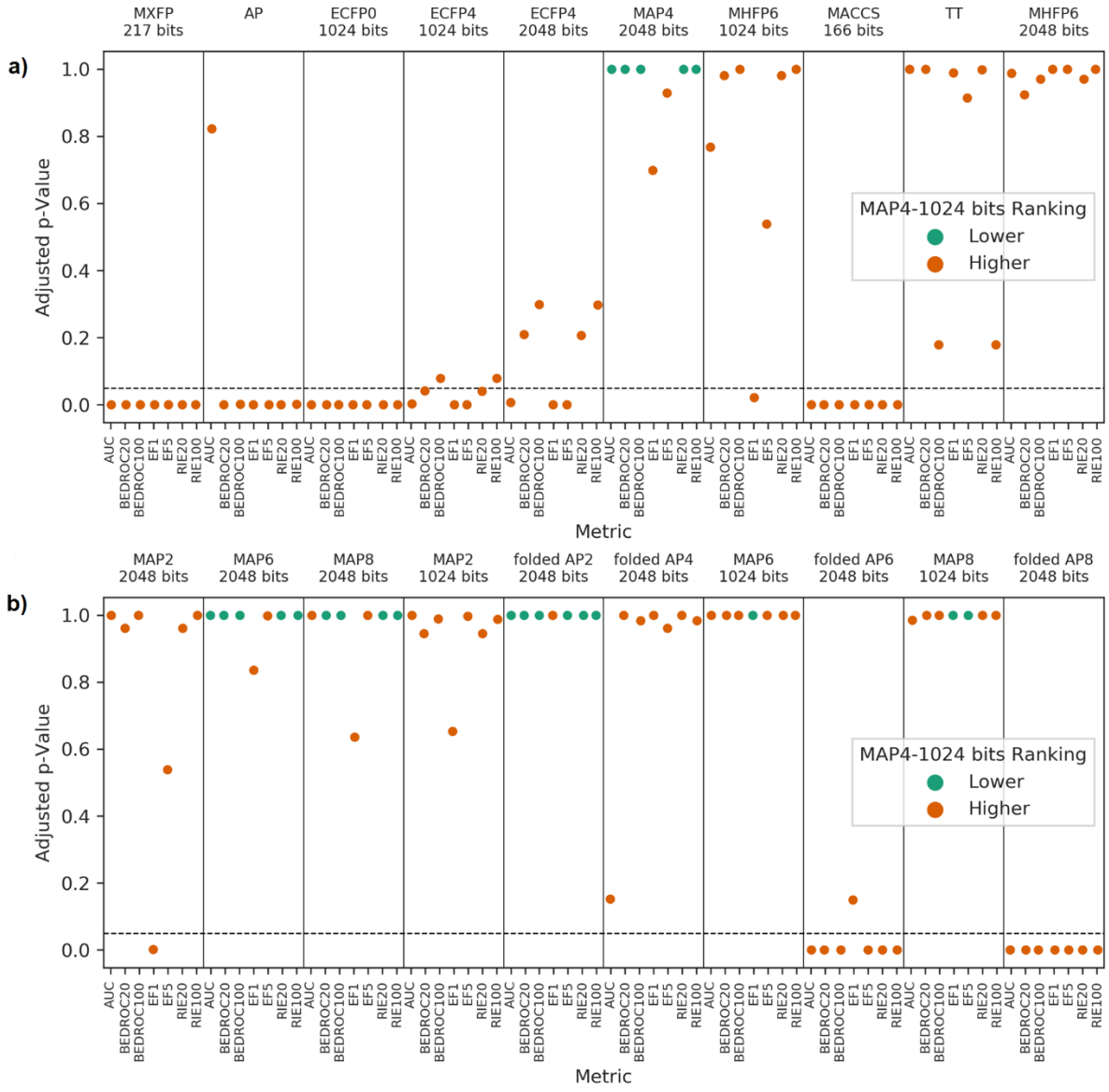


Figure 40. a) Relative ranking and p-values of and MAP4-1024 compared to MXFP-217, AP, ECFP0-1024, ECFP4-1024, ECFP4-2048, MAP4-2048, MHFP6-1024, MACCS-166, TT, and MHFP6 in the Riniker fingerprint benchmark when using only the DUD, MUV, and ChEMBL datasets. b) Relative ranking and p-values of and MAP4-1024 compared to MAP2-2048, MAP6-2048, MAP8-2048, MAP2-1024, foldedAP2-2048, foldedAP4-2048, MAP6-1024, foldedAP6-2048, MAP8-1024, and foldedAP6-2048 in the Riniker fingerprint benchmark when using only the DUD, MUV, and ChEMBL datasets. Orange color corresponds to MHFP6 being ranked higher than the other fingerprint, while green color indicates a lower ranking. P-values below 0.05 (dashed horizontal line) indicate significance.

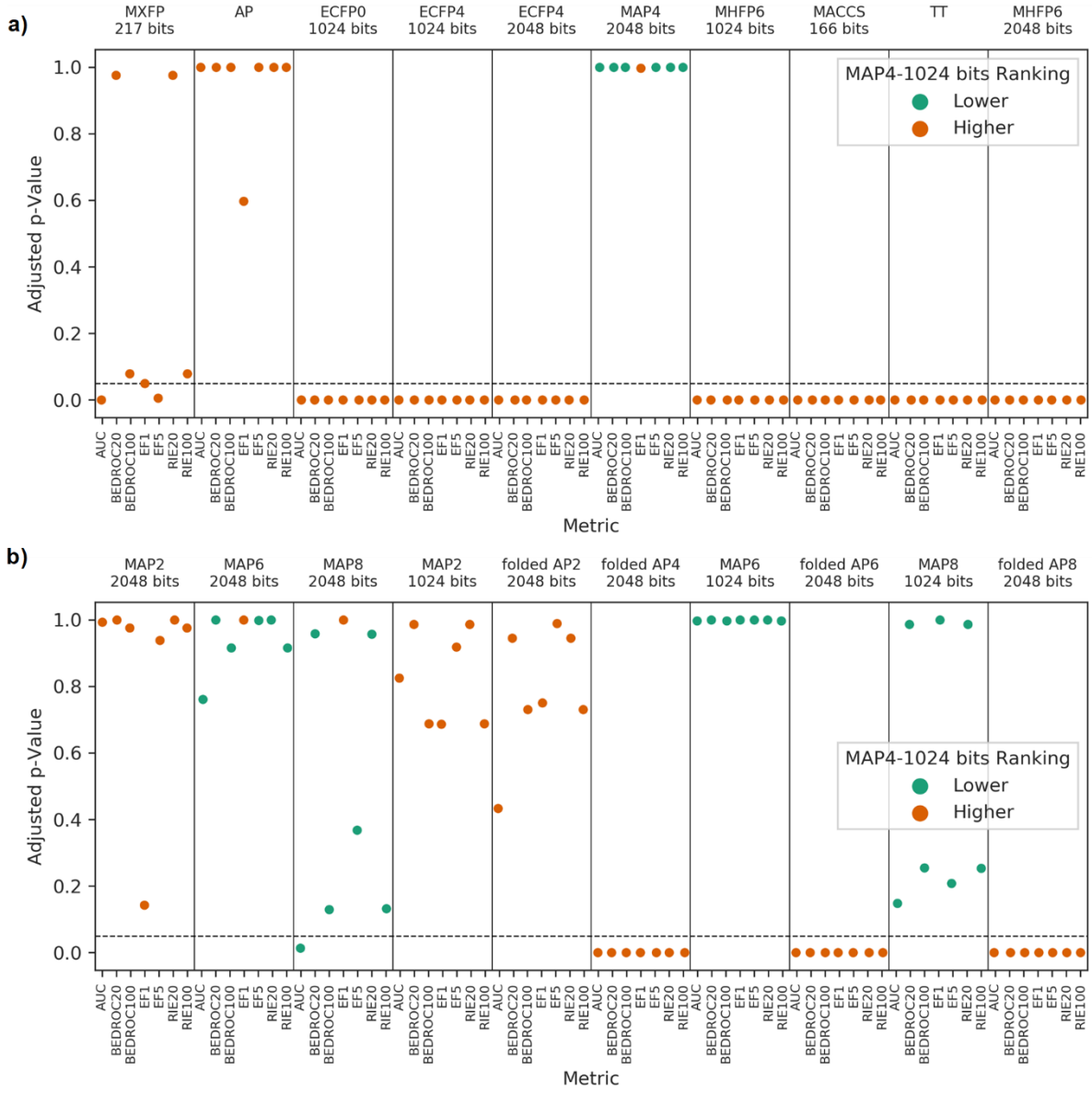


Figure 41. a) Relative ranking and p-values of MAP4-1024 compared to MXFP-217, AP, ECFP0-1024, ECFP4-1024, ECFP4-2048, MAP4-2048, MHFP6-1024, MACCS-166, TT, and MHFP6 in the Riniker fingerprint benchmark when using only the peptide datasets. b) Relative ranking and p-values of and MAP4-1024 compared to MAP2-2048, MAP6-2048, MAP8-2048, MAP2-1024, foldedAP2-2048, foldedAP4-2048, MAP6-1024, foldedAP6-2048, MAP8-1024, and foldedAP6-2048 in the Riniker fingerprint benchmark when using only the peptide datasets. Orange color corresponds to MHFP6 being ranked higher than the other fingerprint, while green color indicates a lower ranking. P-values below 0.05 (dashed horizontal line) indicate significance.

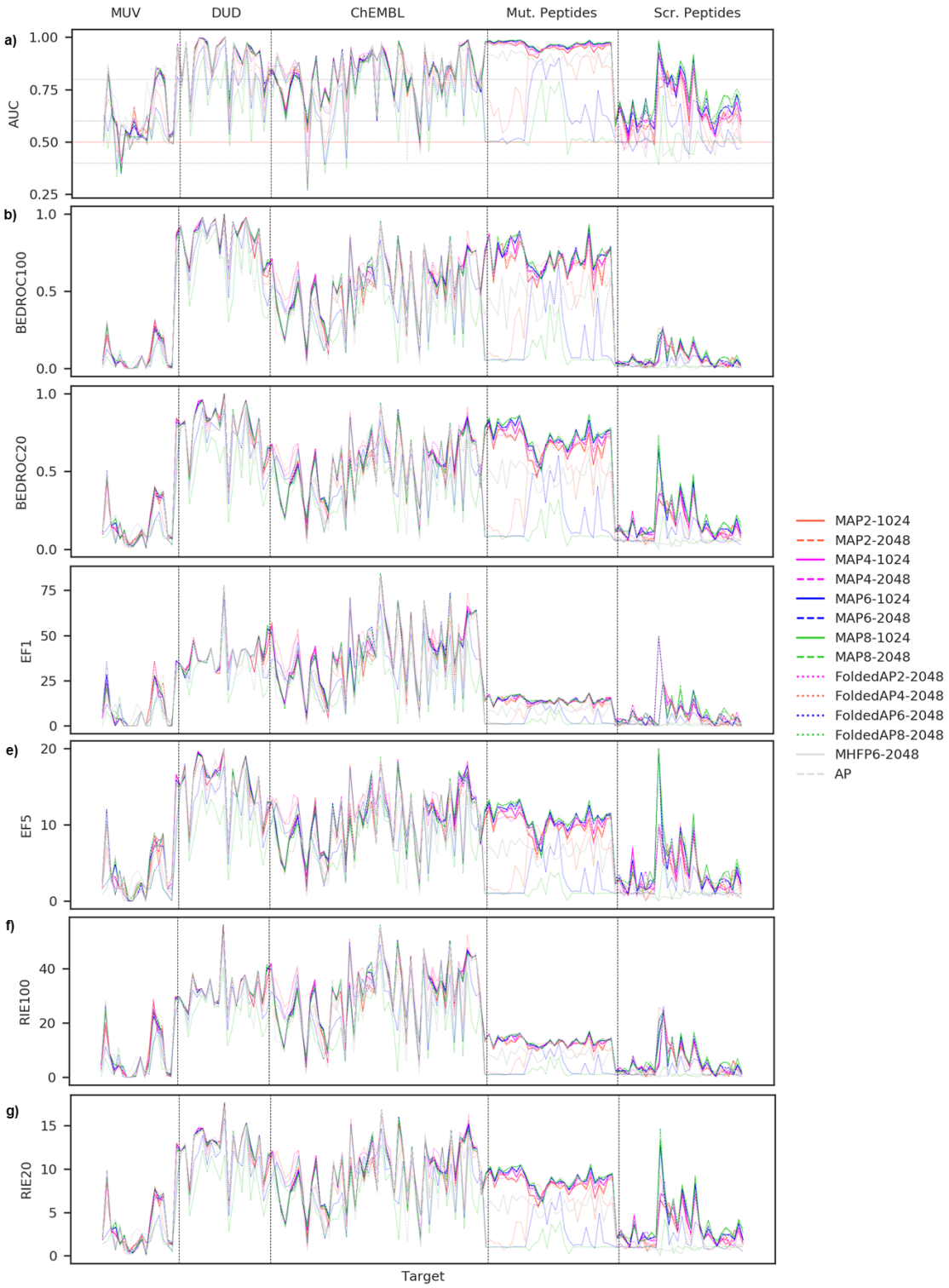
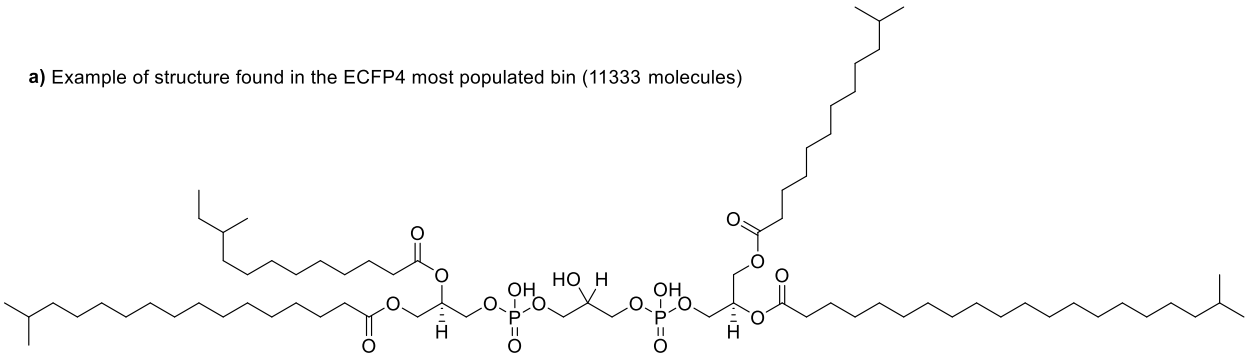


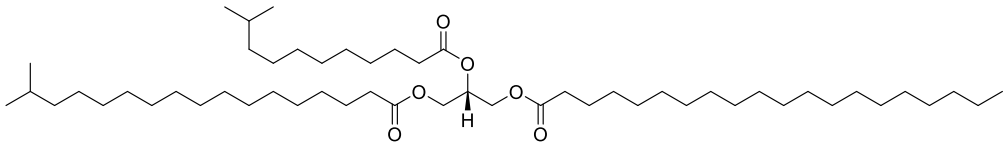
Figure 42. AUC (a), BEDROC100 (b) and 20 (c), EF1 (d) and 5 (e), RIE100 (f) and 20 (g) of AP2 (orange), AP4 (magenta), AP6 (blue), AP8 (green), in their 2014-dimensions (solid) and 2048-dimensions (dashed) minhashed implementation (MAPs), and in their 2048-dimensions folded (dotted) implementation (foldedAPs). MHFP6 (solid) and AP (dashed) are reported in gray.

a) Example of structure found in the ECFP4 most populated bin (11333 molecules)



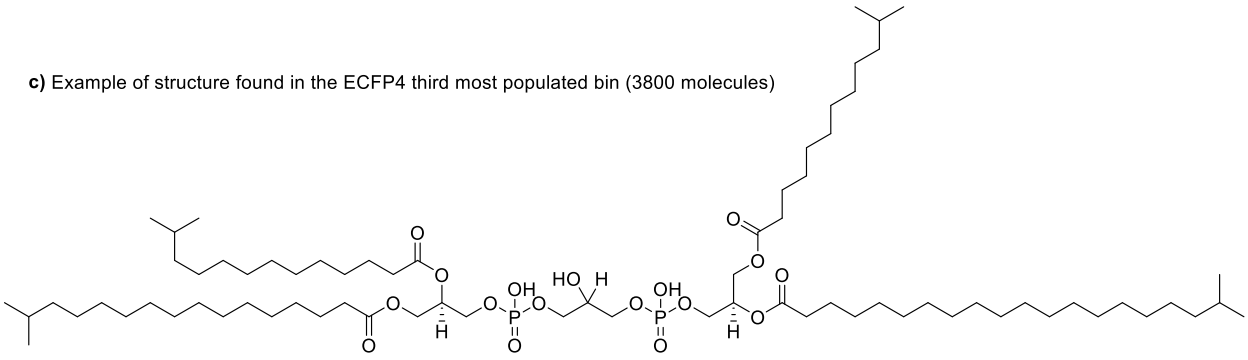
HMDB0072882

b) Example of structure found in the ECFP4 second most populated bin (5694 molecules)



HMDB0108013

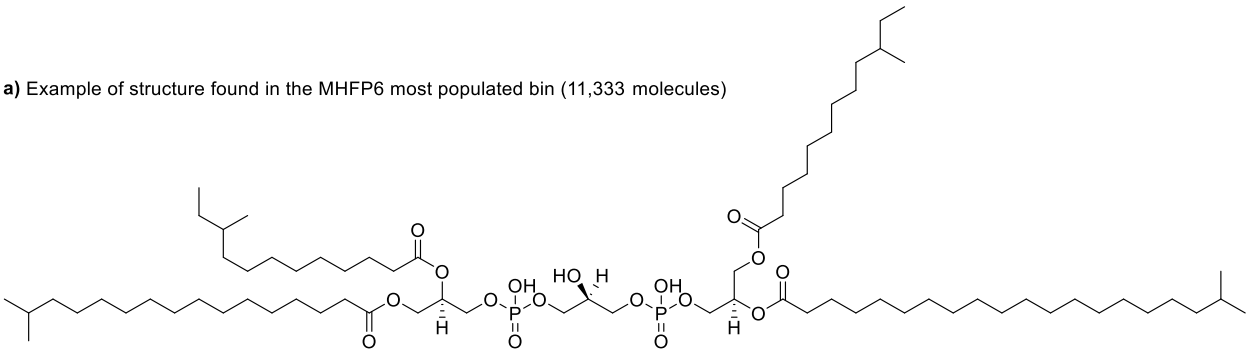
c) Example of structure found in the ECFP4 third most populated bin (3800 molecules)



HMDB72885

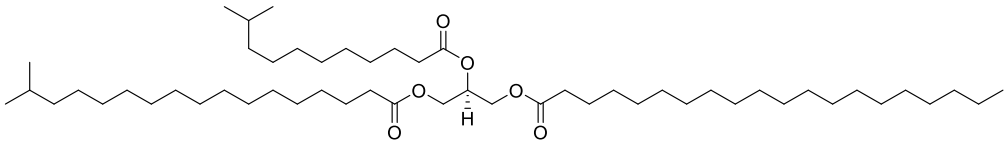
Figure 43. Examples of structures found in the most (a), second-most (b) and third most (c) populated fingerprint value bins in the Metabolome ECFP4-1024 chemical space. The total amount of molecule present in the three fingerprint value bins is reported.

a) Example of structure found in the MHFP6 most populated bin (11,333 molecules)



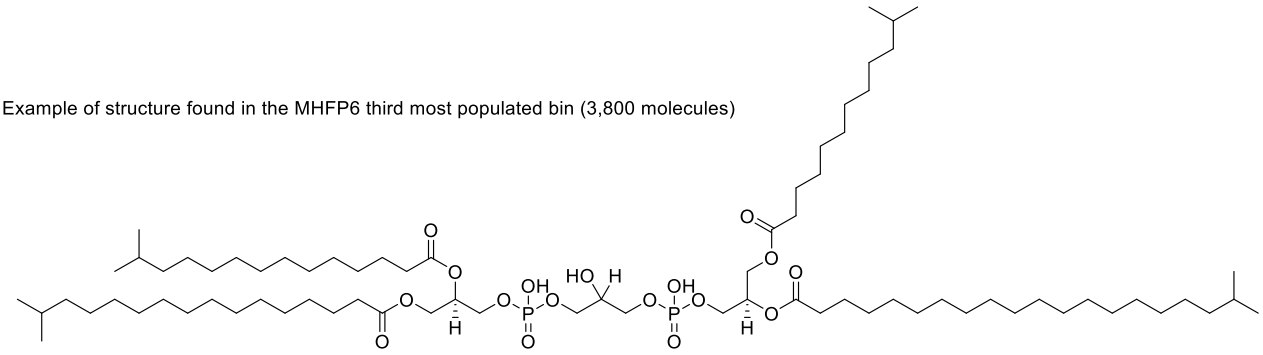
HMDB0072884

b) Example of structure found in the MHFP6 second most populated bin (5,694 molecules)



HMDB0062814

c) Example of structure found in the MHFP6 third most populated bin (3,800 molecules)



HMDB0072895

Figure 44. Examples of structures found in the most (a), second-most (b) and third most (c) populated fingerprint value bins in the Metabolome MHFP6-1024 chemical space. The total amount of molecule present in the three fingerprint value bins is reported.

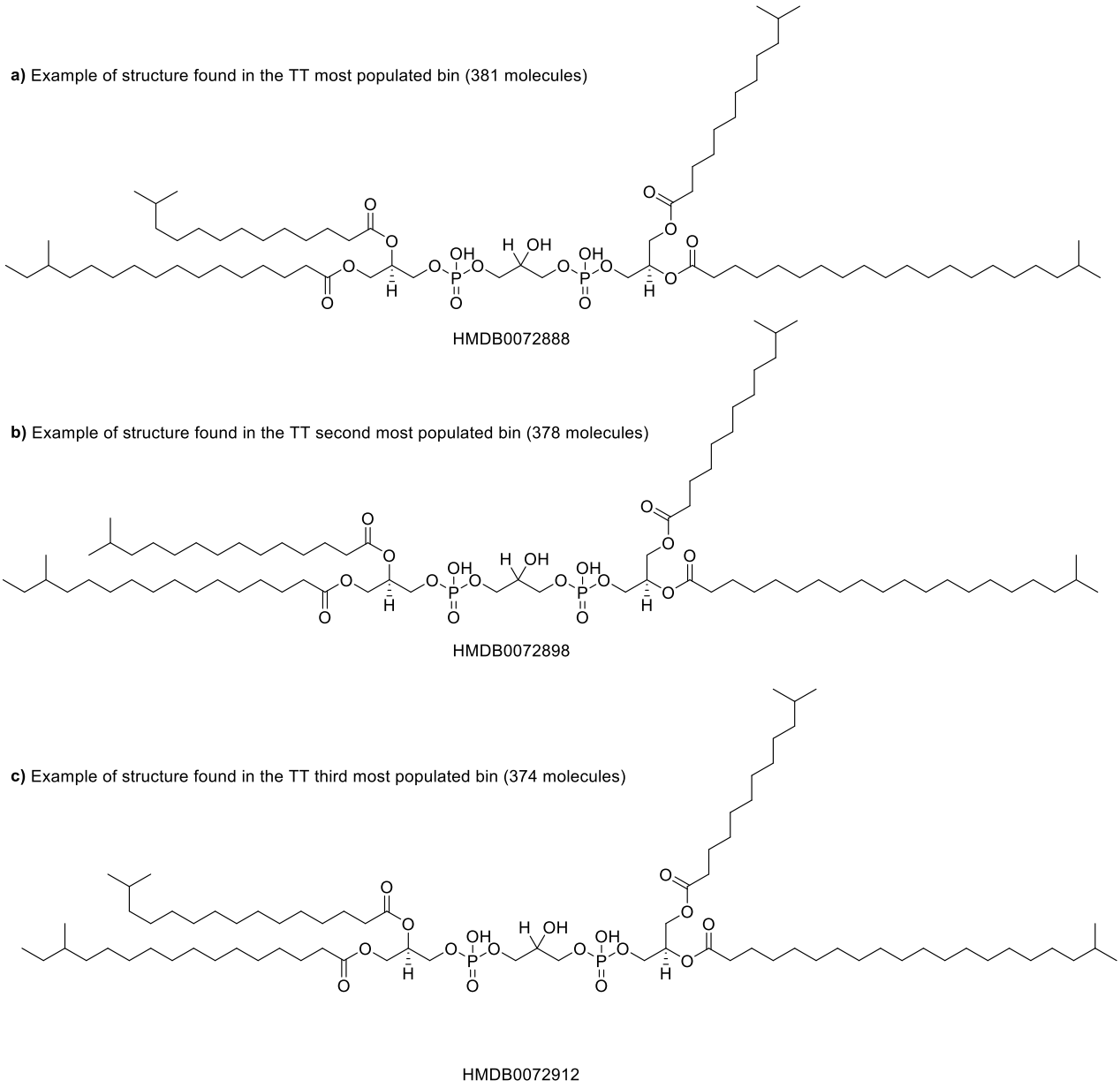


Figure 45. Examples of structures found in the most (a), second-most (b) and third most (c) populated fingerprint value bins in the Metabolome TT chemical space. The total amount of molecule present in the three fingerprint value bins is reported.

Appendix C – Supplementary Tables and Figures for Chapter Four

Note C.1: Table 23

Table 23. MAP4 SVMN classification of known non-microbial natural products.

Natural Product	Origin[a)]	MAP4 SVM pred.[b)] fungal, bacterial	Training set NN	JD from NN[c)]
Salidroside	Plant	**0.75**, 0.25	NPA016219	0.75
Prostacyclin	Animal	**0.94**, 0.06	Shorghumoic acid (NPA005601)	0.78
Serricorole	Animal	0.16, **0.84**	6-deoxyerythronolide B (NPA004018)	0.81
Cholesterol	Animal	**0.93**, 0.07	Micaceol (NPA018196)	0.37
Farnesol	Plant, animal	**0.92**, 0.08	Trans-beta-Farnesene (NPA013150)	0.64
Menthol	Plant	**0.96**, 0.04	(+)-7-Hydroxymenthol (NPA012557)	0.50
Conotoxin MVIIA	Animal	0.13, **0.87**	Siamycin II (NPA020589)	0.75

a) Natural product origin. b) Predicted origin: fungal or bacterial. c) Approximated Jaccard Distance (JD, see methods for details) from the training set NN.

Note C.2: Figures 49-54

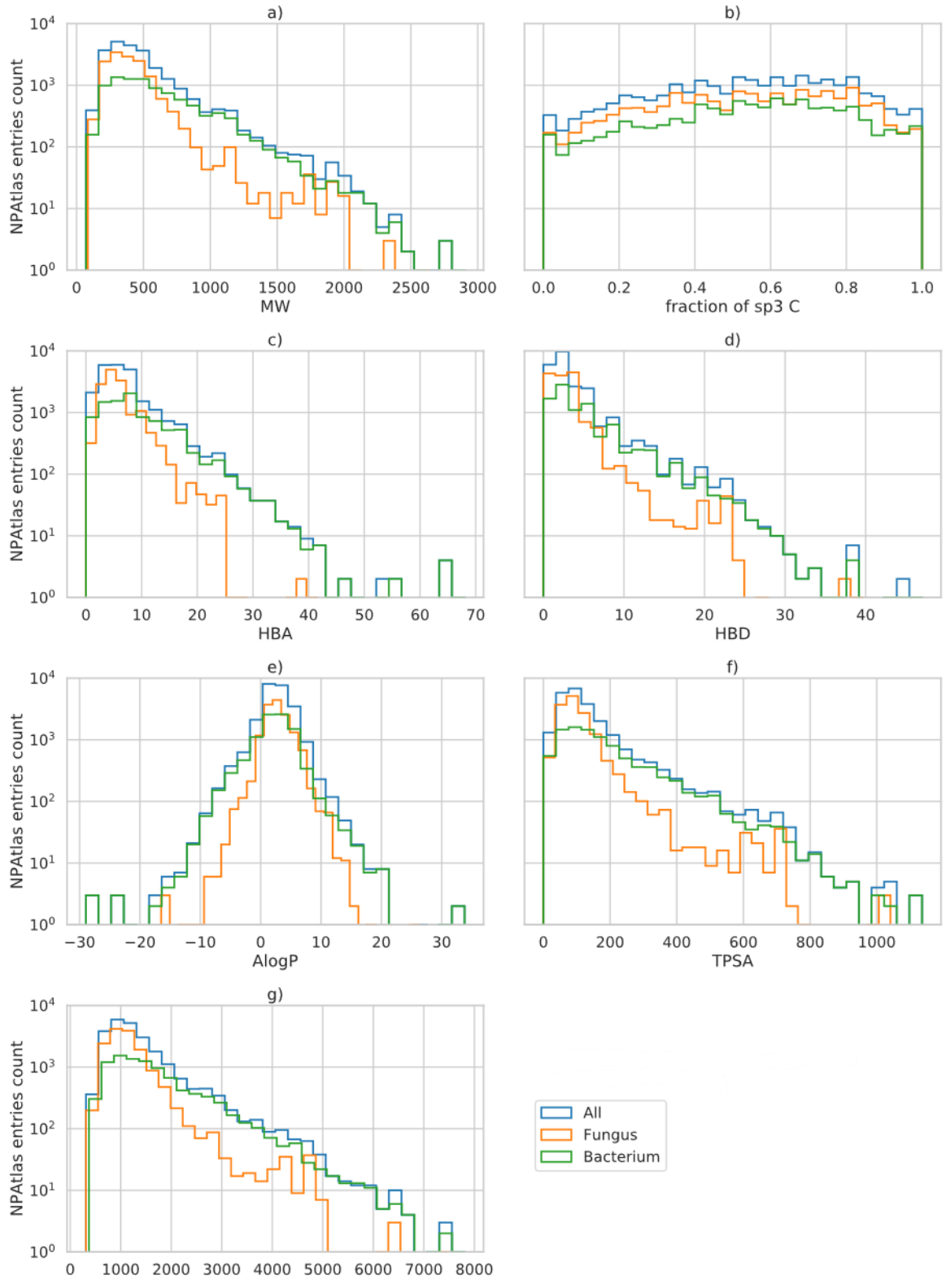


Figure 46. Physico-chemical properties distribution across the NPAtlas entries are shown in blue. The distribution within compounds of fungal and bacterial origin are also reported, in orange and green, respectively.

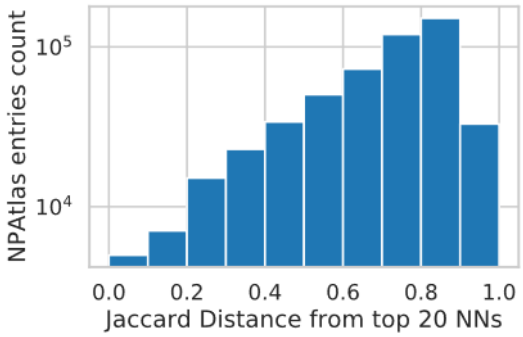


Figure 47. Approximated Jaccard distance from the top 20 NNs of all NPAtlas entries.

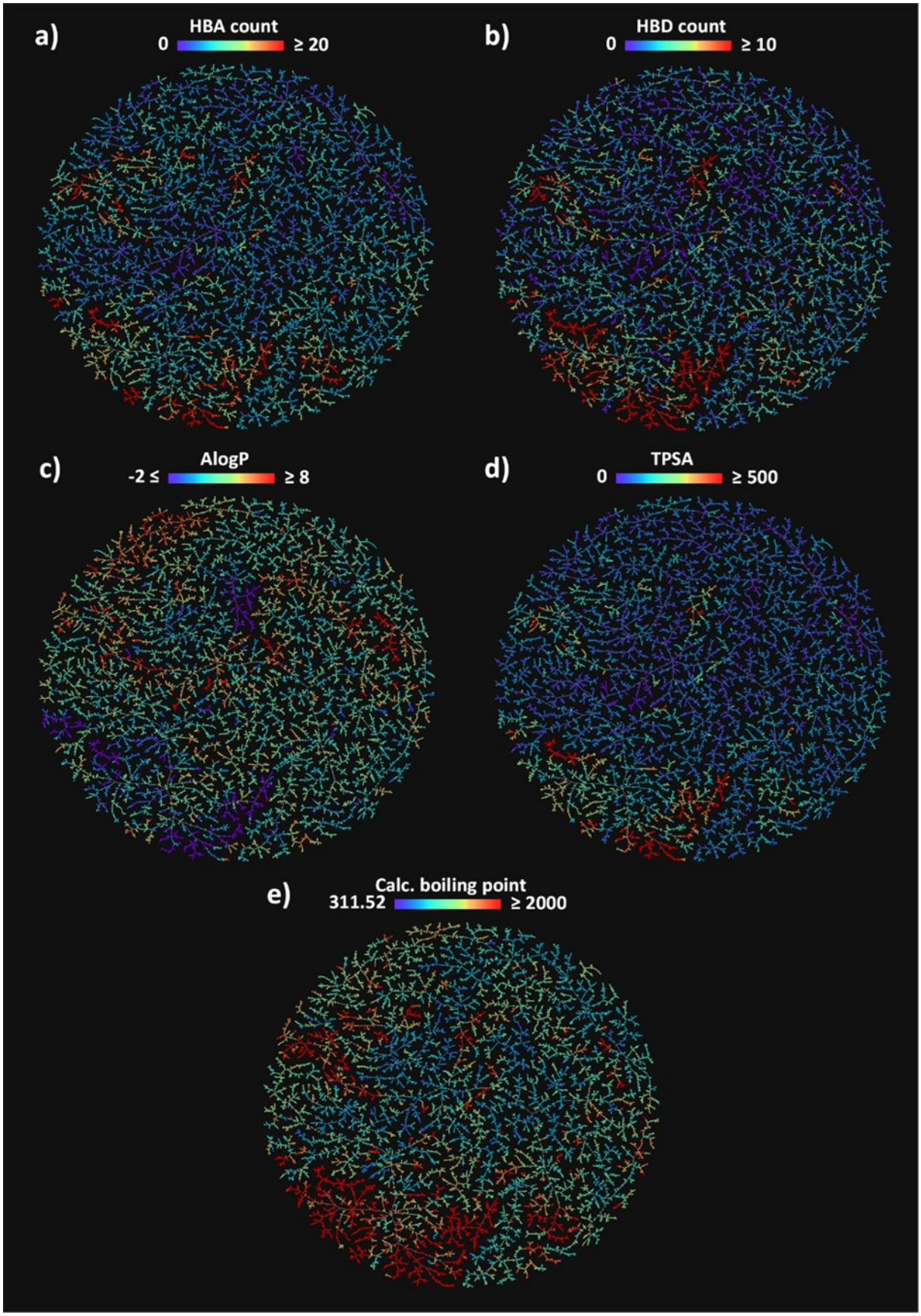


Figure 48. MAP4 TMAP of NPAtlas colored with the available continuous properties (a) HBA, (b) HBD, (c) AlogP, (d) TPSA, and (e) calculated boiling point.

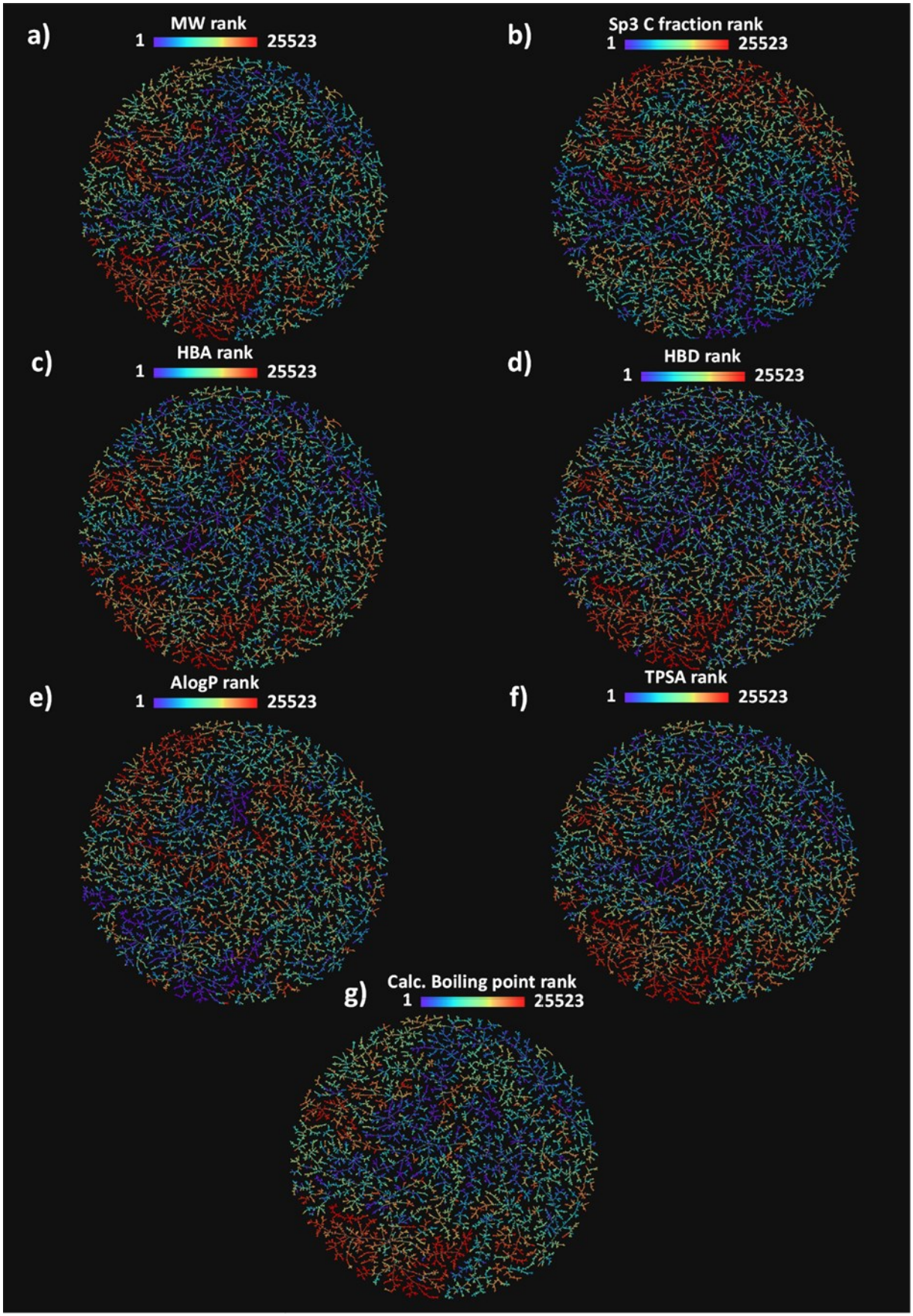


Figure 49. MAP4 TMAP of NPAtlas colored with the ranked continuous properties (a) MW, (b) fsp3C, (c) HBA, (d) HBD, (e) AlogP, (f) TPSA, (g) and calculated boiling point.

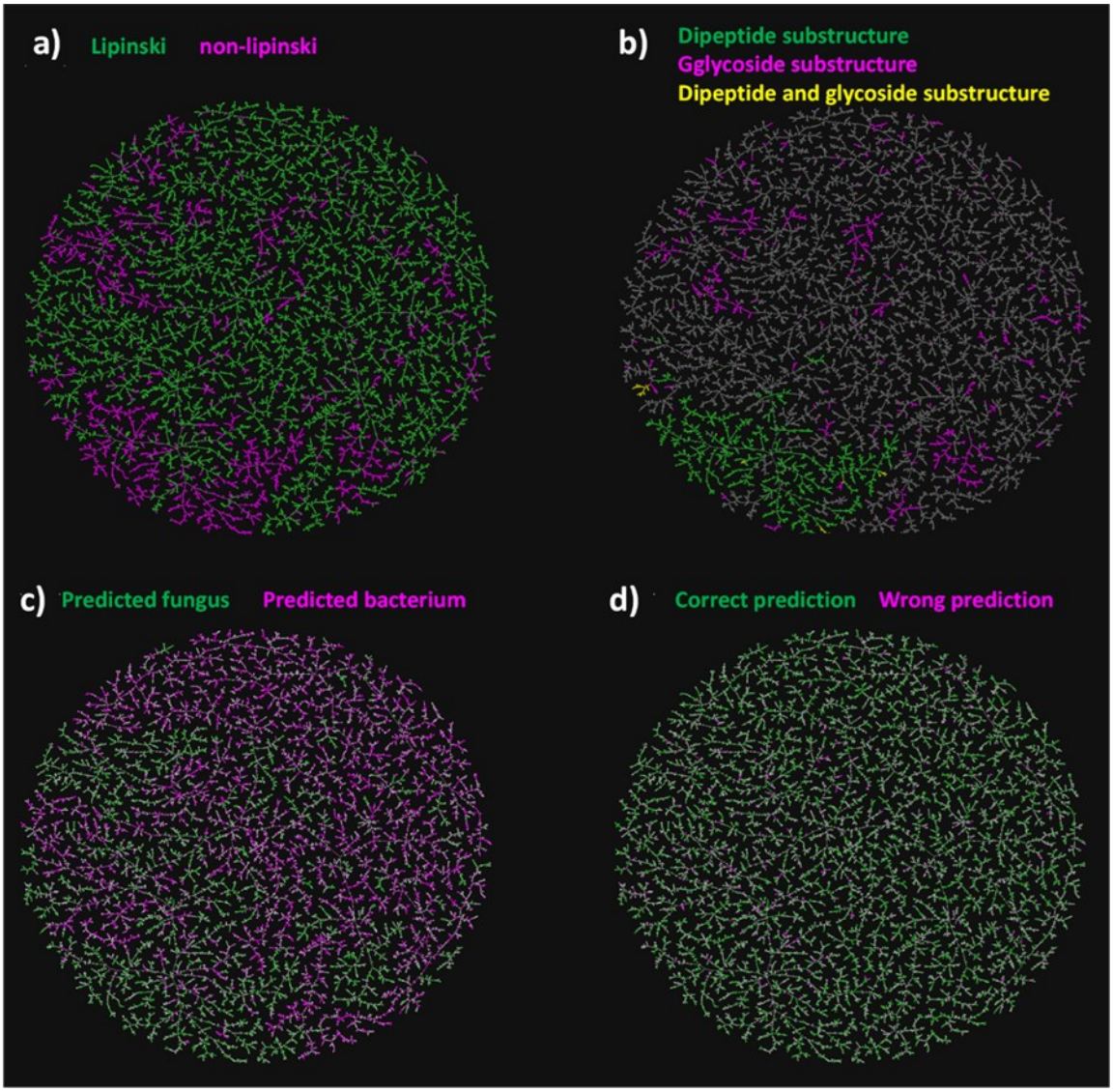


Figure 50. MAP4 TMAP of NPAtlas colored with categorical properties. (a) Lipinski classification. (b) Presence of glycoside and/or dipeptide substructures. (c) MAP4 SMV prediction. (d) MAP4 SVM performance.

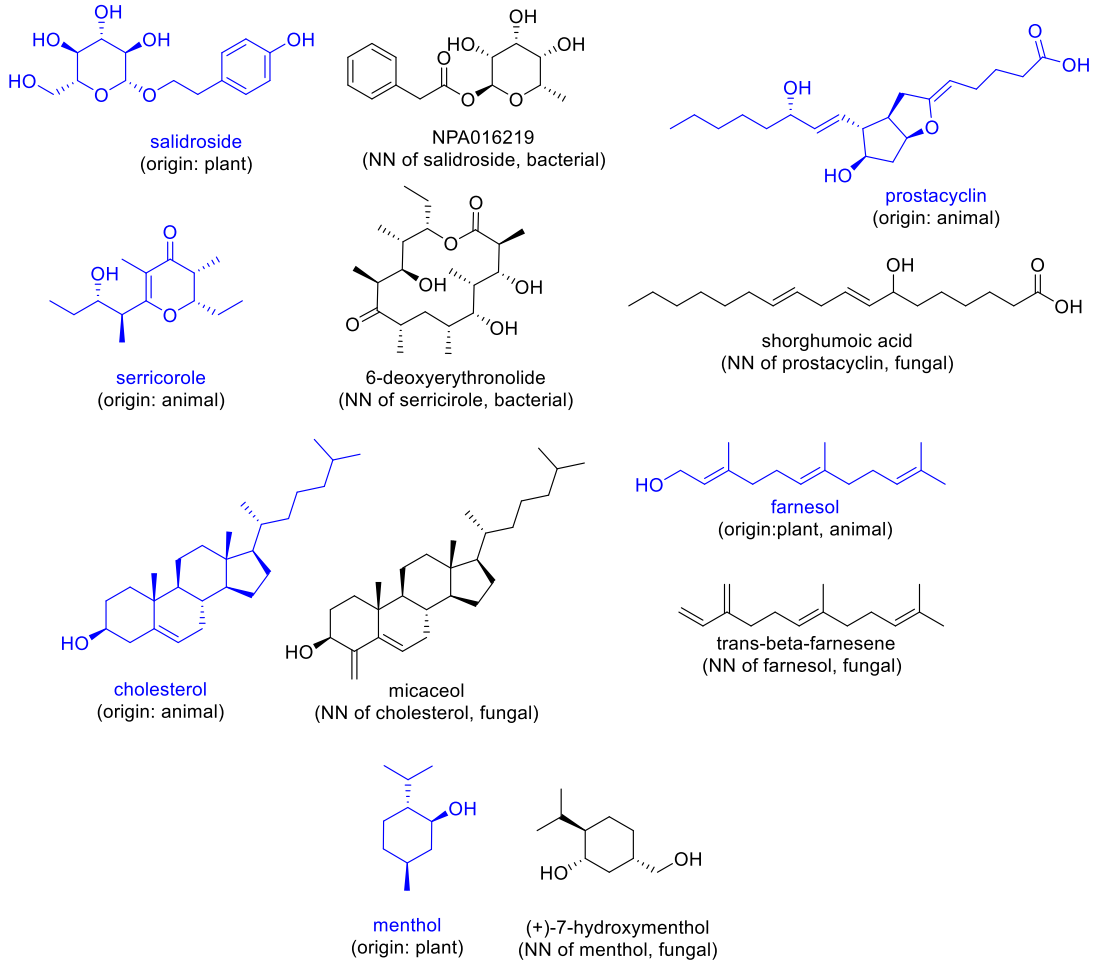
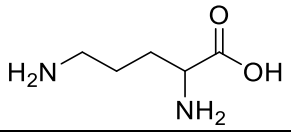
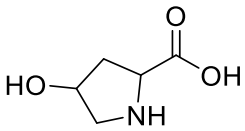
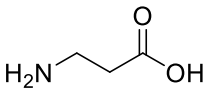
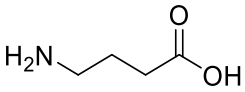
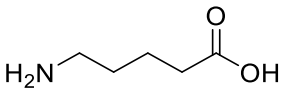
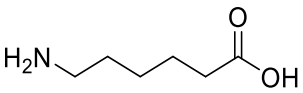
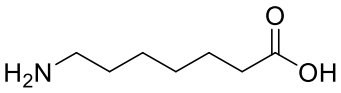
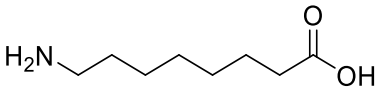
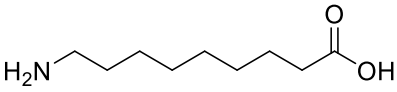
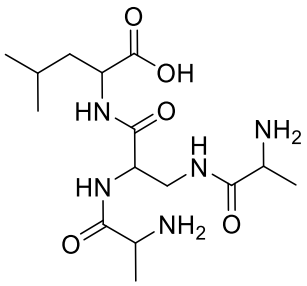


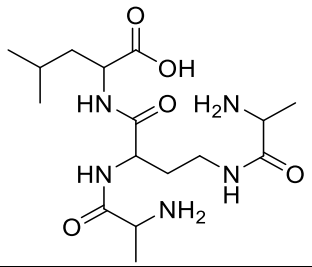
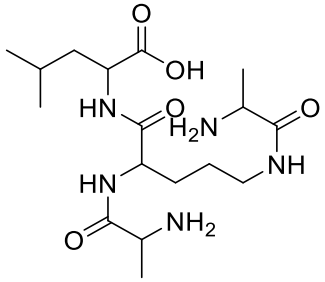
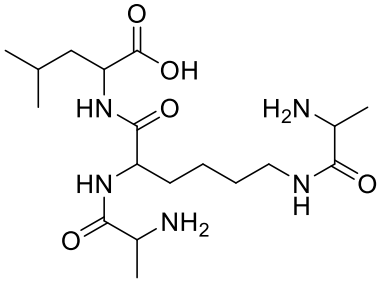
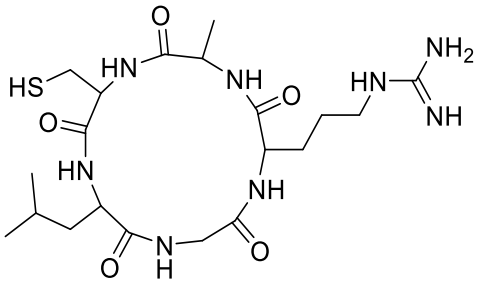
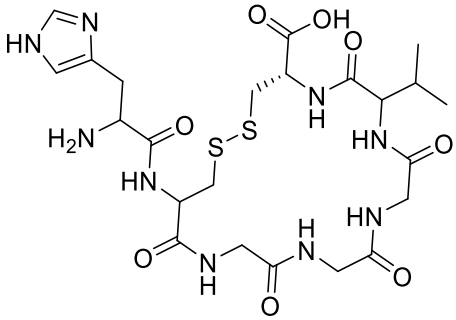
Figure 51. Salidroside, prostacyclin, serricorole, cholesterol, farnesol, menthol, conotoxin MVIIA, and their training set NN. For a better visualization, all query structures are shown in blue.

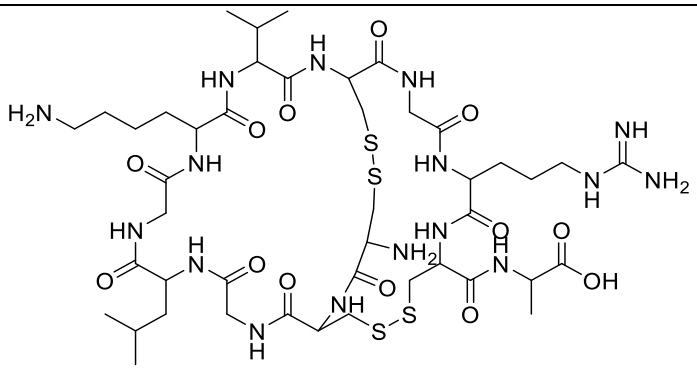
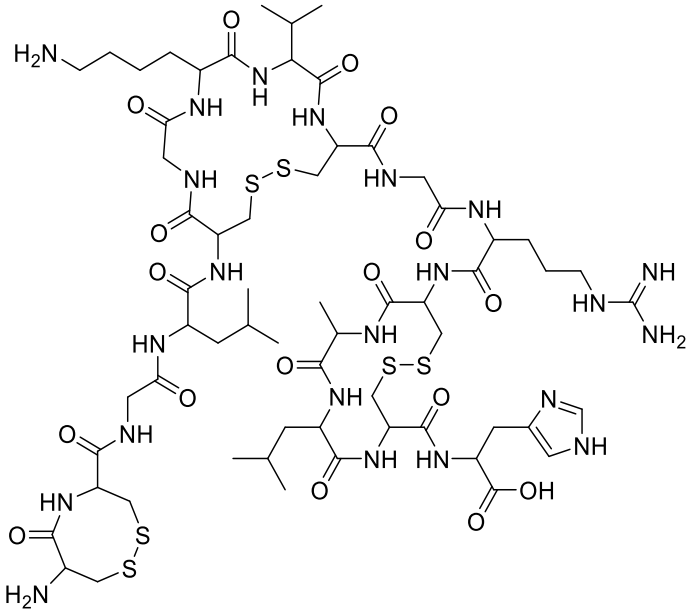
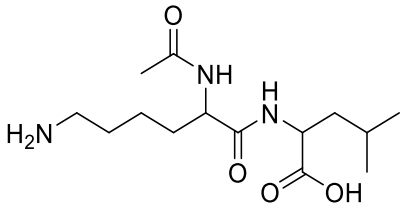
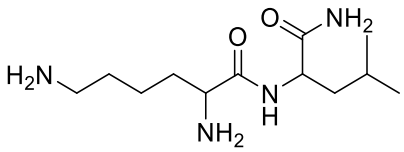
Appendix D – Supplementary Tables and Figures for Chapter Seven

Note D.1: Tables 24-26

Table 24. PDGA recognized symbols and their correspondent internal character.

Symbol[a)]	id[b)]	Description	Example
Orn	O	Ornithine	
Hyp	Z	Hydroxyproline	
bAla	!	Beta-alanine	
Gaba	?	Gamma-aminobutyric acid	
a5a	=	Delta-aminopentanoic acid	
a6a	%	Epsilon-aminohexanoic acid	
a7a	$	Zeta-aminoheptanoic acid	
a8a	@	Eta-aminooctanoic acid	
a9a	#	Theta-aminononaanoic acid	
Dap	1	2,3-diaminopropionic acid as branching unit	Ala-Dap-Leu 
Dab	2	2,4-diaminobutyric acid as branching unit	Ala-Dab-Leu

			
BOrn	3	Ornithine as branching unit	Ala-BOrn-Leu 
BLys	4	Lysine as branching unit	Ala-BLys-Leu 
cy	X	Amide bond head-to-tail cyclization. It is always placed at the beginning (left, N terminus) of the sequence.	Cy-Arg-Ala-Cys-Leu-Gly 
Cys1	Ä	First pair of cyclizes cysteines, Always in pair, never next to each other.	His-Cys1-Gly-Gly-Gly-Val-Cys1 
Cys2	Ö	Second pair of cyclizes cysteines. They are always present in pair, never next to	Cys1-Cys2-Gly-Leu-Gly-Lys-Val-Cys1-Gly-Arg-Cys2-Ala

		each other, present only if Cys1 is already part of the sequence.	
Cys3	Ü	Third pair of cyclizes cysteines. They are always present in pair, never next to each other, present only if Cys1 and Cys2 are already part of the sequence.	Cys1-Cys1-Gly-Leu-Cys2-Gly-Lys-Val-Cys2-Gly-Arg-Cys3-Ala-Leu-Cys3-His 
Ac	&	N-terminus acetylation. It is always placed at the beginning (N-terminus, left) of the sequence	Ac-Lys-Leu 
NH2	+	C-terminus amide. It is always placed at the end (C-terminus, right) of the sequence	Lys-Leu-NH2 

[a] PDGA input. [b] Internally used characters. [c] Lower case.

Table 25. SMARTS HBA/HBD and charge assignment.

Description	SMARTS	HBA/HBD [a) b)]	Charge [b)]
Aliphatic carbon	[C]	0	0
Nitrogen	[#7]	1	0
Tertiary nitrogen	[NX3;H0]	0	0
Aliphatic oxygen	[O]	2	0
Thiol	[SH1X2]	1	0
Hydroxyl at aliphatic carbon	[$([OHX2]-[CX4])]	3,0	0,0
Tyrosine	N[CX4H1]([CH2X4][cX3]1[cX3H][cX3H][cX3]([OHX2,OH0X1-])[cX3H][cX3H]1)[CX3]=[OX1]	1,0,0,0,0,0,0,3,0,0,0,2	0,0,0,0,0,0,0,0,0,0,0,0
Proline	N1[CX4H]([CH2][CH2][CH2]1)[CX3](=[OX1])	0,0,0,0,0,0,2	0,0,0,0,0,0,0
Histidine	N[CX4H]([CH2X4][#6X3]1:[$([#7X3H+,#7X2H0+0]:[#6X3H]:[#7X3H]),$([#7X3H])]:[#6X3H]:[$([#7X3H+,#7X2H0+0]:[#6X3H]:[#7X3H]),$([#7X3H])]:[#6X3H]1)[CX3](=[OX1])	1,0,0,0,2,0,1,0,0,2	0,0,0,0,0,0,0,0,0,0
Charged nitrogen	[$([NH2X3]-[CX4]),$([N]=[CX3])]	1,0	1,0
Carbonyl	[$([O]=[C])]	2,0	0,0
Carboxyl	[OH,O-]-[C](=O)	2,0,2	-1,0,0
Ether	[OX2]([CX4])[CX4]	2,0,0	0,0,0
Phenol	[OH1X2]-[c]	3,0	0,0

[a)] 0 = no hydrogen donor or acceptor site; 1 = donor site; 2 = acceptor site; 3= donor and acceptor site. [b)] When more values are present, they refer to the SMARTS atom in the correspondent position.

Table 26. Input parameters and excluded building blocks (bb).

Target	N	M	G	Treshold CBD	Topology	Excluded bb [a)]	Time limit
Indolicidin	50	1	0.8	300	linear	Hyp, Orn, bAla, Gaba, a5a, a6a, a7a, a8a, a9a, Ac	24 h
Indolicidin SEQSIM	50	1	0.8	5	linear	Hyp, Orn, bAla, Gaba, a5a, a6a, a7a, a8a, a9a, Ac	6 h
Tyrocidine A	50	1	0.8	300	cyclic	bAla, Gaba, a5a, a6a, a7a, a8a, a9a, Ac	24 h
ω-conotoxin-MVIIA	50	1	0.8	300	polycyclic	Hyp, Orn, bAla, Gaba, a5a, a6a, a7a, a8a, a9a, Ac	72 h
G3KL	50	1	0.8	300	dendrimer	Hyp, Orn, bAla, Gaba, a5a, a6a, a7a, a8a, a9a, Ac	48 h
Acetyl-CoA	50	1	0.8	300	linear	Ac	24 h
Epothilone A	50	1	0.8	300	cyclic	Ac	24 h
Cholic Acid	50	1	0.8	300	cyclic	Ac	24 h
α-cyclodextrin	50	1	0.8	300	cyclic	Ac	24 h
PAMAM dendrimer	50	1	0.8	300	dendrimer	Ac	24 h

[a)] For a definition of the mentioned building blocks refer to **Table 24**.

Note D.2: Figure 55

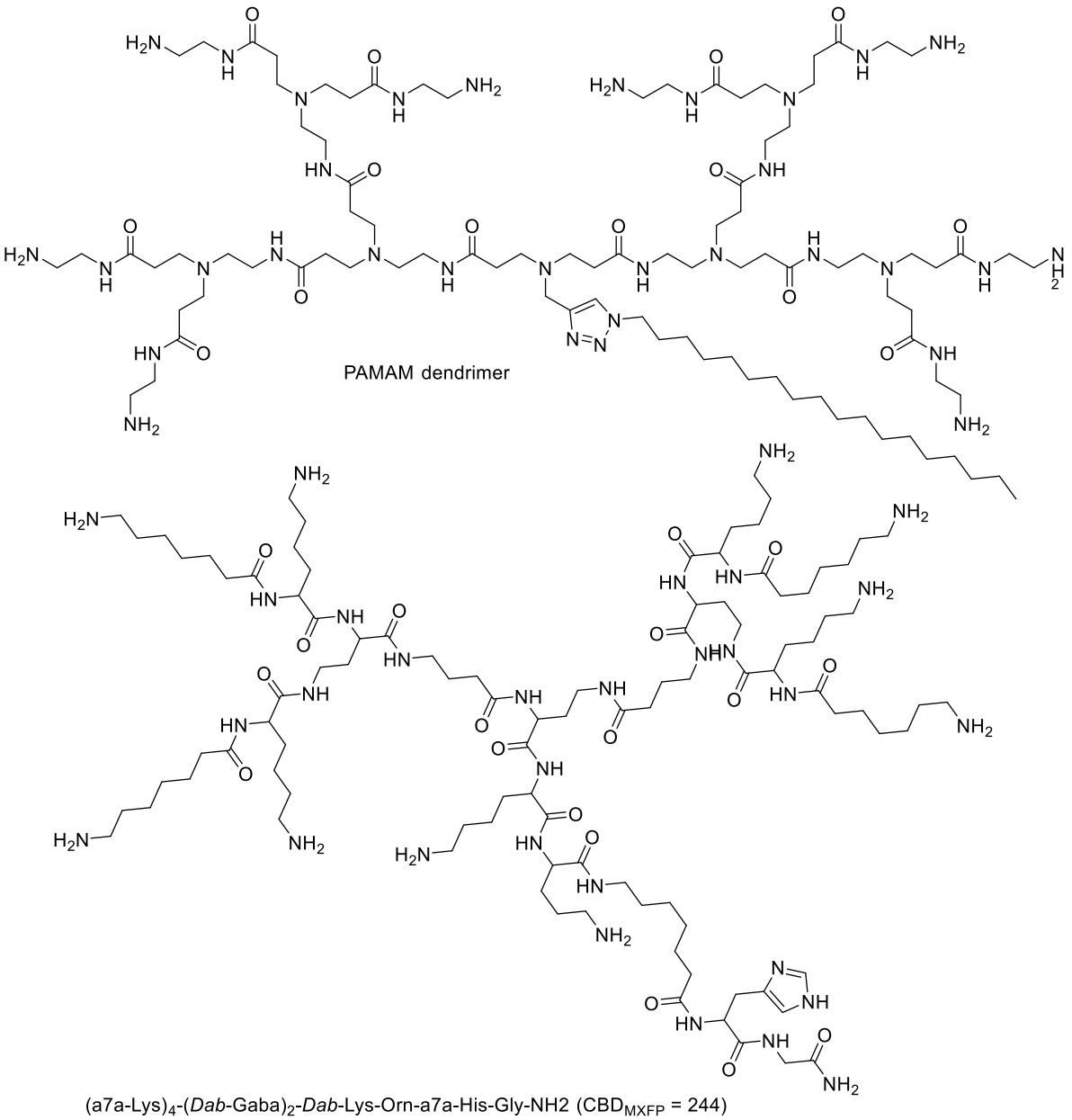


Figure 52. Structures of the non-peptidic PAMAM target and its best analog.

Appendix E – Supplementary Tables and Figures Chapter Eight

Note E.1: Tables 27,28

Table 27. Performance on the test of the NB, RF, SVM, RNN, and RNN with scrambled labels (RNN scr.) models for the AMP activity (act.) and hemolysis (hem.) classification tasks.

Classifier	ROC AUC	Accuracy [a)]	Precision [a)]	Recall [a)]	F1 score [a)]	MCC [a)]
NB act.	0.55	0.55	0.59	0.32	0.42	0.11
SVM act.	0.75	0.68	0.68	0.68	0.68	0.36
RF act.	0.81	0.71	0.70	0.75	0.73	0.44
RNN act.	**0.84**	**0.76**	**0.74**	**0.80**	**0.77**	**0.53**
RNN scr. act.	0.51	0.49	0.35	0.03	0.05	-0.06
NB hem.	0.58	0.56	0.48	**0.76**	0.59	0.19
SVM hem.	0.69	0.73	0.72	0.58	0.65	0.44
RF hem.	0.80	**0.77**	**0.81**	0.60	0.69	**0.53**
RNN hem.	**0.87**	**0.76**	0.70	**0.76**	**0.73**	**0.52**
RNN scr. hem.	0.45	0.61	0.41	0.05	0.10	0.01

a) The probabilistic prediction values were converted into binary classification values using a threshold of 0.5. The best values and the selected classifiers are reported in bold.

Table 28. Classifiers optimization

Classifier	**Hyperparameters optimization** [a)]
SMV AMP activity	C = 0.1, 1, **10**, 100 γ = 0.1, 0.01, **0.001**
RF AMP activity	maximum depth = **10**, 30, 50, 70, 90, None no. estimators = 10, 100, 250, 500, 750, **1000**, 1500, 2000
RNN AMP activity	embedding dimensions = 2, 21, 42, **100** GRU dimensions = 50, 100, 200, 300, **400** no. layers = 1, **2**, 3 epoch = [1, 2, 3, ..., 150]; best epoch = **38**
RNN AMP activity scrambled labels	embedding dimensions **2**, 21, 42, 100 GRU dimensions = **50**, 100, 200, 300, 400 no. layers = **1**, 2, 3 epoch = [1, 2, 3, ..., 150]; best epoch = **1**
SMV Hemolysis	C = 0.1, **1**, 10, 100 γ = 0.1, 0.01, **0.001**
RF Hemolysis	maximum depth = **10**, 30, 50, 70, 90, None no. estimators = 10, 100, 250, **500**, 750, 1000, 1500, 2000
RNN Hemolysis	embedding dimensions = 2, 21, 42, **100** GRU dimensions = 50, 100, 200, 300, **400** no. layers = **1**, 2, 3 epoch = [1, 2, 3, ..., 150]; best epoch = **95**
RNN Hemolysis scrambled labels	embedding dimensions **2**, 21, 42, 100 GRU dimensions = **50**, 100, 200, 300, 400 no. layers = 1, **2**, 3 epoch = [1, 2, 3, ..., 150]; best epoch = **150**

a) The selected hyperparameters are highlighted in bold. All hyperparameters that have not been discussed have been used in their default values.

Note E.2: Figures 56-60

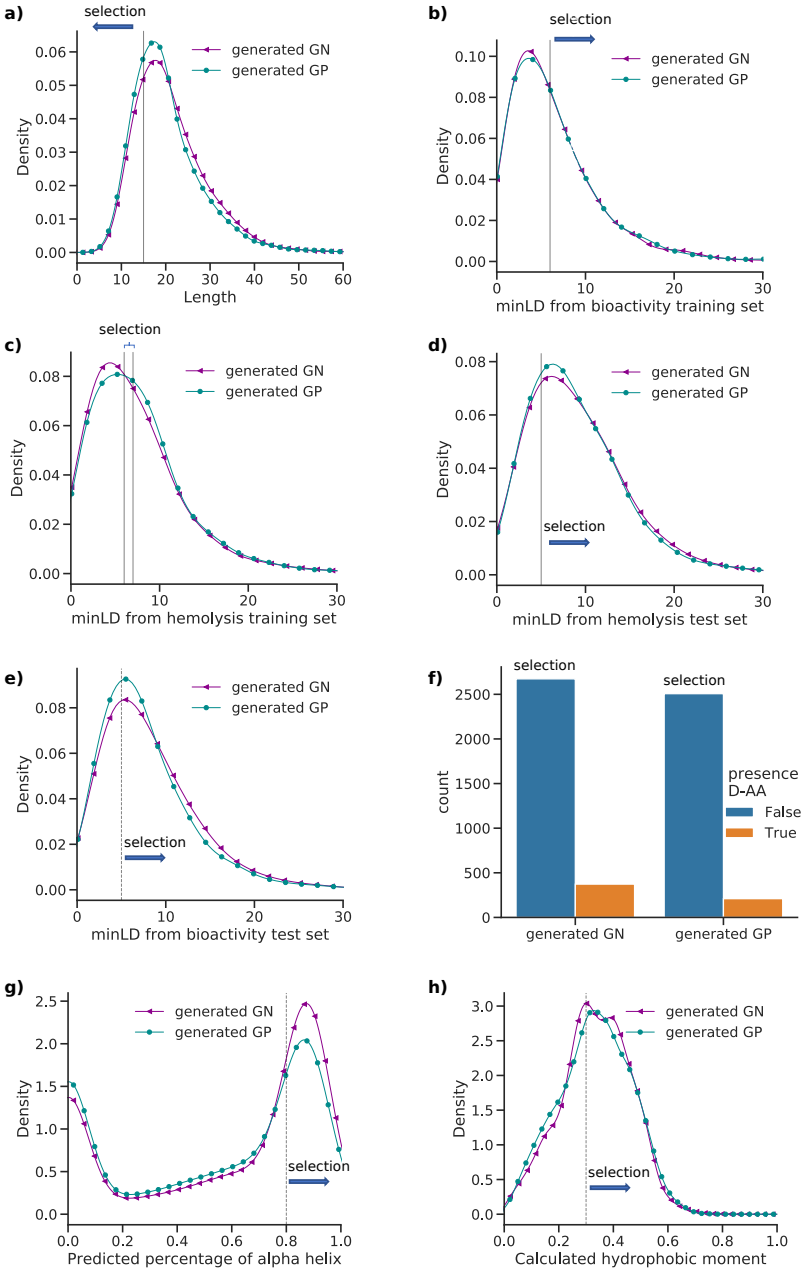


Figure 53. Properties distribution and filters. (a) length, (b-e) minimum Levenshtein distance (minLD) from training and test sets, (f) presence of D-amino acids (D-AA), (g, h) Amphiphilic helix estimation, of the 3,046 predicted active and non-hemolytic sequences derived from the model fined tuned for *A. baumannii* and *P. aeruginosa* (Generated GN) and the 2,717 predicted active and non-hemolytic sequences derived from the model finetuned for *S. aureus*. (a-d) Solid vertical lines indicate that the threshold values were included. (e, g, h) Dashed vertical lines indicated that the threshold lines were excluded.

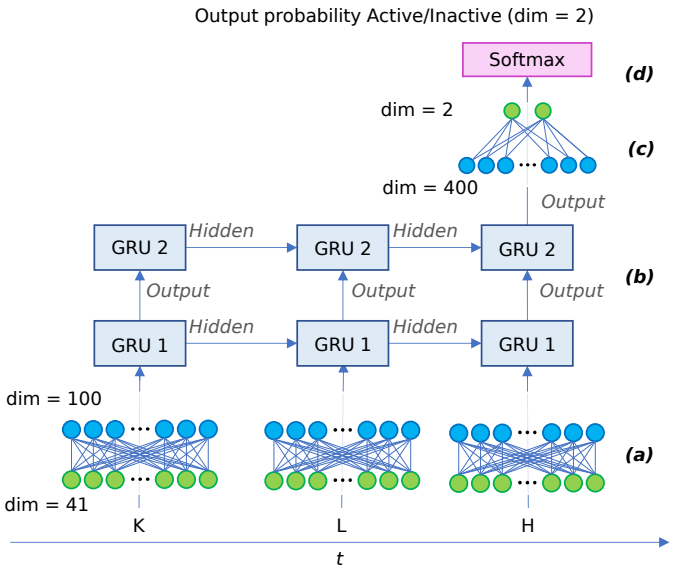


Figure 54. RNN AMP activity classifier architecture. The tokenized and "one-hot" encoded sequences enter an 100 dimensions (dim) embedding layer (a), then they are processed through two layers of 400 dimensions GRU cells (b), and finally, a linear transformation layer shapes the last GRU output into two dimensions (c), followed by a softmax function that normalizes it into a probability (d). The architecture of the hemolytic classifier differs only by having one layer of GRU cells.

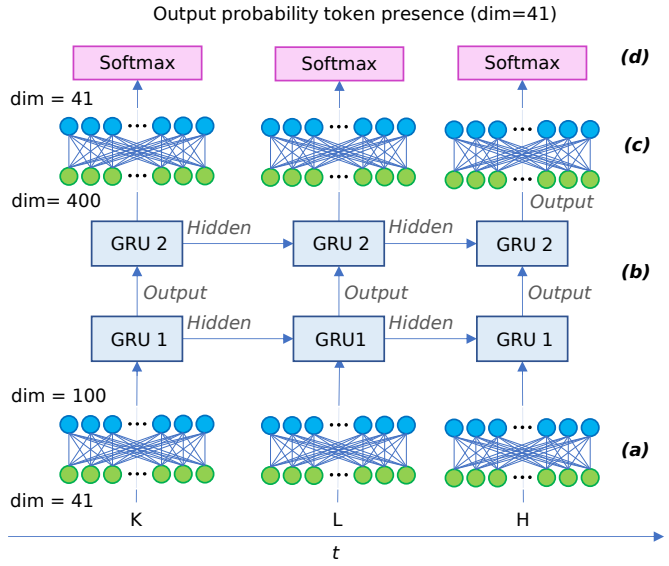


Figure 55. RNN generative models architecture. The tokenized and "one-hot" encoded sequences enter an 100 dimensions (dim) embedding layer (a), then they are processed through two layers of 400 dimensions GRU cells (b), and finally, a linear transformation layer shapes the last GRU output into 41 dimensions (c), followed by a softmax function that normalizes it into a probability (d). The architecture of the hemolytic classifier differs only by having one layer of GRU cells.

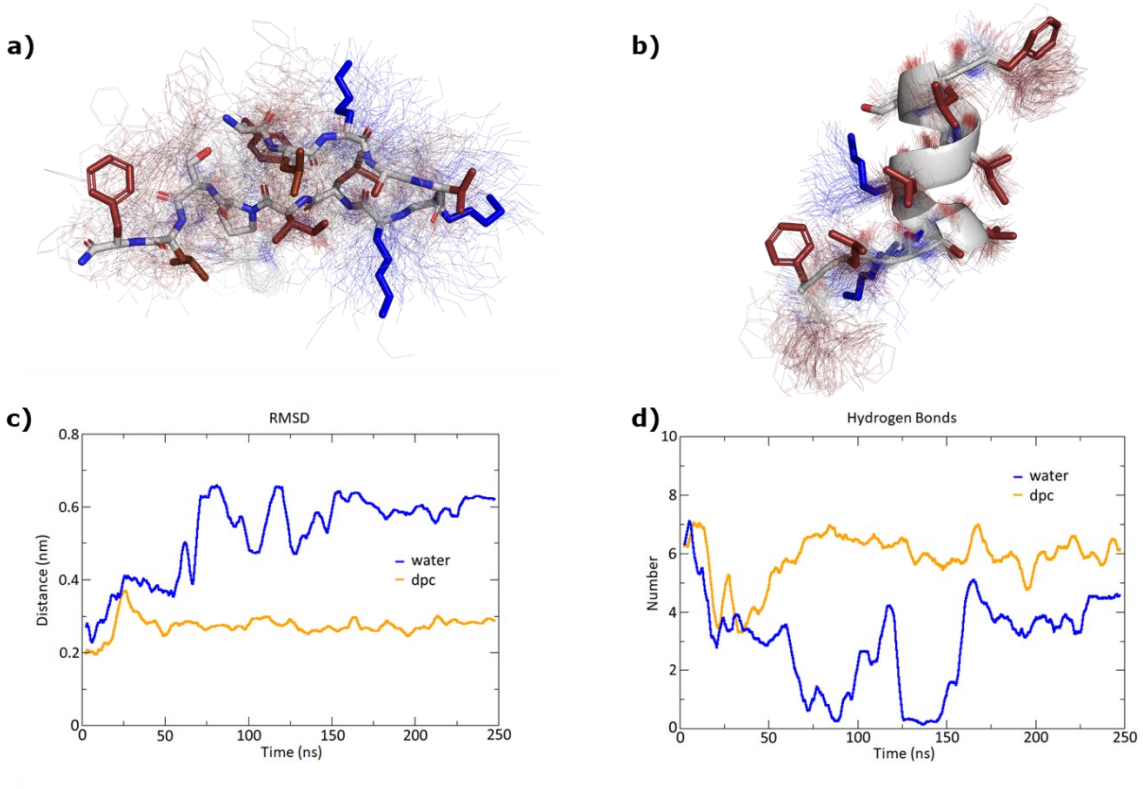


Figure 56. MD simulations of **GP1** in water and in presence of a DPC micelle over 250 ns using GROMACS. (**a**) Average structure (stick model) in water over 100 structures sampled over the last 100 ns (thin lines). Hydrophobic side chains are colored in red and cationic side chains in blue. (**b**) Average structure (cartoon model for backbone and stick model for side chains) with DPC micelle over 100 structures sampled over the last 100 ns (thin lines). (**c**) RMSD (root mean square deviation) of the peptide backbone atoms relative to the starting α-helical conformation. (**d**) Number of intramolecular hydrogen bonds. The DPC micelle was omitted for clarity.

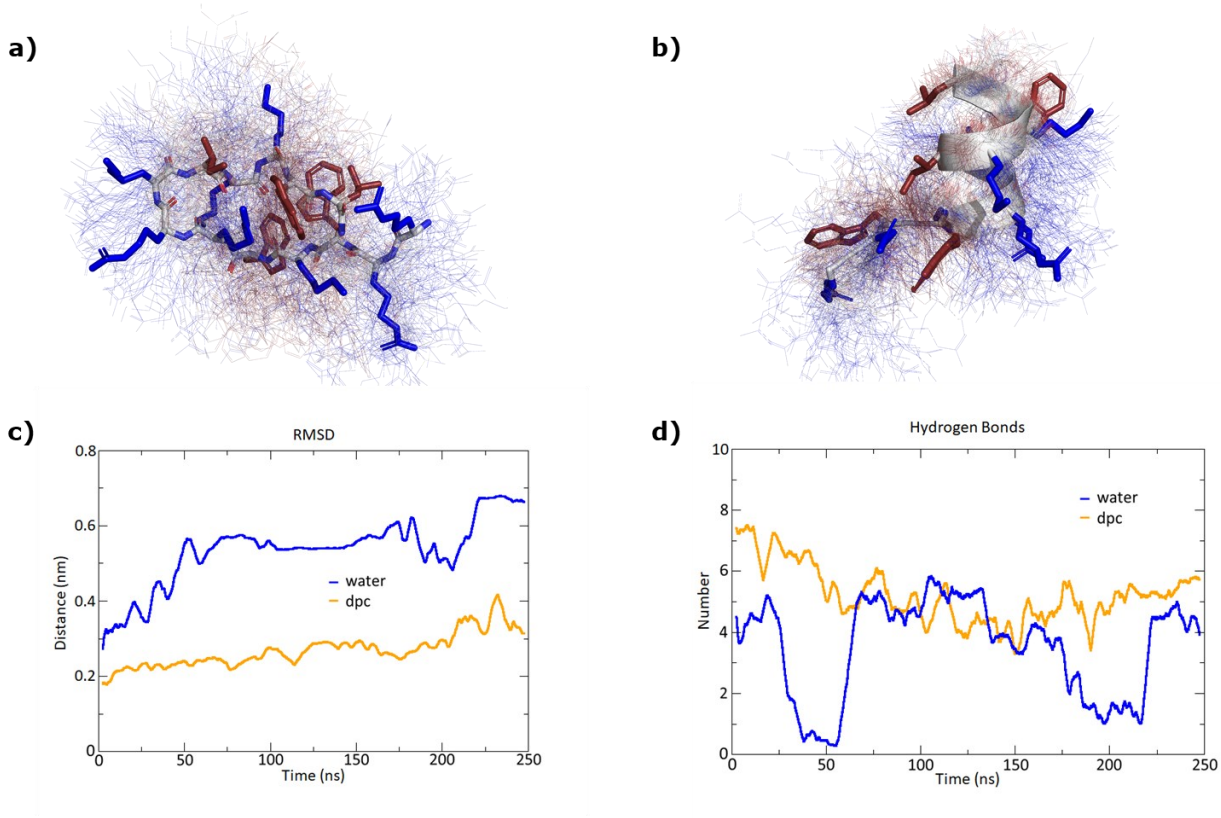


Figure 57. MD simulations of **GN2** in water and in presence of a DPC micelle over 250 ns using GROMACS. (**a**) Average structure (stick model) in water over 100 structures sampled over the last 100 ns (thin lines). Hydrophobic side chains are colored in red and cationic side chains in blue. (**b**) Average structure (cartoon model for backbone and stick model for side chains) with DPC micelle over 100 structures sampled over the last 100 ns (thin lines). (**c**) RMSD (root mean square deviation) of the peptide backbone atoms relative to the starting α-helical conformation. (**d**) Number of intramolecular hydrogen bonds. The DPC micelle was omitted for clarity.

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Capecchi Alice

Registration Number: 17-134-230

Study program: Doctorate in Chemistry and Molecular Sciences

Bachelor ☐ Master ☐ Dissertation ☑

Title of the thesis: Cheminformatics Tools to Explore the Chemical Space of Peptides and Natural Products

Supervisor: Prof. Dr. Jean-Louis Reymond

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of September 5th, 1996 and Article 69 of the University Statute of June 7th, 2011 is authorized to revoke the doctoral degree awarded on the basis of this thesis.
For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with theses submitted by others.

Bern,

Place/Date

Signature

ALICE CAPECCHI

PROFILE

Sympathetic, passionate, team player.
PhD candidate in Cheminformatics.
After a master thesis in Medicinal Chemistry and a two-years graduate program in AstraZeneca, I have pursued my passion in research and started a PhD in Cheminformatics at the University of Bern.
I am currently writing my PhD thesis, and it is time to look for an exciting job opportunity that will allow me to contribute to drug discovery with the professional and personal expertise gather in these years.

CONTACT

EMAIL:alice.capecchi@outlook.it
PHONE: +41 (0)77 510 55 14
LINKEDIN: linkedin.com/in/alicecapecchi
GITHUB: github.com/alicecapecchi
GOOGLE SCHOLAR: scholar.google.com/citations?user=4miJLsAAAAAJ

LANGUAGES

Italian – native
English – proficient
German – beginner (A2)

ACTIVITIES AND INTERESTS

Great vegetarian food
Craft beer
House plants
Postmodern surreal literature

WORK EXPERIENCE

PhD in Cheminformatics with a Focus on Peptides and Large Molecules
University of Bern, Switzerland
2017 – present (expected October 2021)
Acquired expertise:

- Data analysis and visualization
- Virtual screening
- Machine learning, deep learning
- Programming languages and toolkits: Python, Bash, RDKit, PyTorch, scikit-learn, and Pandas
- Scientific project management (resulted in seven first authored publications).

AstraZeneca Trainee – innovative Medicines
AstraZeneca R&D, Gothenburg, Sweden
2015 – 2017
Two years trainee program structured in three projects within different departments:

- Computational Chemistry – library design and Pharmacophore Screening
- Automation Chemistry – synthesis on plate of two libraries previously designed during the placement in Comp. Chem. Along with follow up and analysis of the assays results
- Structure and Biophysics – crystallization and analysis of protein structures

EDUCATION

Master's Degree in Pharmaceutical Chemistry and Technology
University of Pisa, Italy
2009 – 2014
110 / 110 cum laude; University of Pisa, Italy
Thesis project: "Synthesis and Computational Study of New Inhibitors of Lactate Dehydrogenase"

FURTHER TRAINING

- Oxford Group development modules: self-understanding, developing of communication skills, presentation and influencing techniques, SMART objectives, recognition of stakeholders
- BSP international "Leadership Skills for Leaders": recruitment and appraisal interviews, team building, management by objectives and conflict training
- Writing for Publication in Chemistry and Biochemistry

KEY PUBLICATIONS

- Machine learning designs non-hemolytic antimicrobial peptides. **A Capecchi**, X Cai, H Personne, T Köhler, C van Delden, J-L Reymond, *Chemical Science*, 2021
- One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. **A Capecchi**, D Probst, J-L Reymond, *J. Cheminformatics*, 2020