Computational Strategies for the Data-Driven Discovery of Antimicrobial Peptides

Inaugural dissertation

of the Faculty of Science,

University of Bern

presented by

Markus Orsi

From Bolzano, Italy

Supervisor of the doctoral thesis:

Prof. Dr. Jean-Louis Reymond

Department of Chemistry, Biochemistry and Pharmaceutical Sciences

Computational Strategies for the Data-Driven Discovery of Antimicrobial Peptides

Inaugural dissertation

of the Faculty of Science,

University of Bern

presented by

Markus Orsi

From Bolzano, Italy

Supervisor of the doctoral thesis:

Prof. Dr. Jean-Louis Reymond

Department of Chemistry, Biochemistry and Pharmaceutical Sciences

Accepted by the Faculty of Science.

Bern, **15th July 2025**

The Dean Prof. Dr. Jean-Louis Reymond

« Don’t Panic. »

Douglas Adams, *The Hitchhiker’s Guide to the Galaxy*

Acknowledgements

I was told the acknowledgements are the one part of a thesis you should enjoy writing. It's also the one part I kept postponing until the very end. But if there's one thing a PhD teaches you, it's how to deliver things right before the deadline.

To Jean-Louis: Thank you for your guidance, trust, and patience over the years. I'm grateful for the freedom you gave me to shape my work and for always making time, even with your impossibly full calendar. You taught me to think independently and, perhaps more importantly, never discouraged me from pursuing ideas that were unconventional, exploratory, and sometimes a little unorthodox. I hope to carry that curiosity forward.

To the Reymond Group and Collaborators: Sandra, Sacha, Alice, Amol, Geo, Elena, Aline, Mario, Hippolyte, Kleni, David, Celine, Etienne, Thierry, Ye, Sam, Jeremie, Basak, Yves, Maedeh, Matheus, Leon, Xiaoling, Austia, Robin, Giulia, Yasien, Jacopo, Alex, Bee Ha, Maria, Andrea, Ziad, Angelo, Çagri, Mirco - *you are the heart of this thesis*.

Thank you for the science, the interruptions at every coffee break and Z'vieri, the questionable memes, the intense lunch table debates, and the collective groans at every Wednesday progress report. Among all the memories, a few stand out especially: Sacha, for being the most opinionated person in the universe, but also incredibly helpful, direct, and blissfully free of sugarcoating. Geo, for being the most iconic member of the group, hands down. Aline, for your food and wine recommendations, always 10/10. Hippolyte, for saving me from ever needing to Google impact factors (you really know them all). Kleni, *per essere stata la mia italiana sostitutiva nei momenti di carestia linguistica, faleminderit*. David, for always being around when I needed a second brain (you had enough for both of us). Céline, for your energy and, of course, the (in)famous meme wall. Etienne, for being my collaborator on my favourite project and my midday sport partner for maintaining sanity. Thierry, for

testing the boundaries of acceptable humour. Lucky for you, I'm not easily offended. Ye, 谢谢你总是记得我的生日，也在很多小事上支持我. Sam, for the chaotic videos, the Aareböötle adventures, and the coolest tattoos I've ever seen. Jeremie, for intense science talks and shady comments that always come out of nowhere. Başak, for the beer, the collaboration, and the humour. Yves - I do not Yves-n need to further elaborate. Your support has meant a lot, in science, in sports, and in life. Maedeh, for our shared taste in too many things, but especially for introducing me to lavashak. Matheus, for radiating calm, clarity, and quiet confidence (and the pão de queijo). Leon, *deine Begeisterung für all deine Hobbies ist einfach extrem ansteckend*. Xiaoling, 谢谢你教我中文，也合作得很愉快. Austia, for bonding over my Lululemon addiction and questionable taste in books and movies. Robin and Jacopo, for the football updates. Giulia, *per la tua energia positiva e travolgente; sei sempre una ventata d'aria fresca (o forse un tornado)*. Alex, for the wide-ranging, always stimulating discussions. Angelo, Çagri, and Mirco, thank you for the collaboration and for the great time we had during our overlap in Bern. And of course, thank you Sandra for shielding us from the labyrinth of university bureaucracy, for solving problems before we even knew they existed, and for always doing it with kindness and calm.

To my family: Special thanks to my parents, Greti and Ivo, for their unwavering support throughout all these years of study (at least the PhD came at no extra cost for you). Thank you for always believing in me and encouraging me to follow my own path. And to the rest of my family, thank you for always being a safe haven and distraction when I'm back home in Südtirol. Josh and Asja, *mi mancate! An die Familie, die ich in den letzten Jahren dazugewonnen habe: Danke für eure Unterstützung und dafür, dass ihr mich mit so großer Selbstverständlichkeit aufgenommen habt.*

To Fabian: You've heard every rant, calmed every panic, and cheered on every small win, all while delivering your own (more frequent) rants with flair. Thank you for being my sounding board, my anchor, and my favourite distraction.

To my friends: I feel incredibly lucky to be surrounded by a strong and steady support system, people who have seen me at my best and worst, and stuck around for all the in-betweens. Paul, Chiara, Ben, Sébastien, thank you for being my everyday anchors, for tolerating my nonsense and rants, and for always keeping things light. You made the long days shorter and the short days better, and I cherish every moment I get to spend with you. Fabian, Calu, Lara (whenever you're not off exploring some corner of the world), *grazie per trovare sempre il tempo di vedermi ogni volta che torno a casa, vi voglio bene.* Julian, Elias, Theo, Heidi, *a wenn's net immer kloppt ins za segn, es isch jeds mol aso als hattn mir ins nia aus di Augn verlorn*. Leander, you recently left us, but your presence hasn't. I carry your memory lightly, but always. To all my friends spread around Switzerland: *äuä,* you truly make this country feel like a second home to me.

To everyone I forgot: Pretend I wrote something poetic and deeply moving about you. It is probably true.

This section wasn't peer-reviewed, but it's written with the same care as any scientific paper. Thank you all for shaping these years into something far more meaningful than just a degree.

General Abstract

Cheminformatics has played a central role in medicinal chemistry, enabling the storage, analysis, and modelling of large volumes of chemical data, particularly for small organic molecules. However, its application to large and structurally complex compounds remains underdeveloped. This thesis addresses that gap by developing and improving computational tools that extend molecular representation and modelling strategies to natural products, modified peptides and macromolecules, which often fall outside the scope of conventional methods.

One part of the thesis focuses on the reimplementation and extension of two molecular fingerprints. The macromolecule extended atom-pair fingerprint (MXFP) was adapted within an open-source framework and applied to the analysis of chemical spaces composed of molecular pairs. Separately, the MinHashed atom-pair fingerprint (MAP4) was extended to encode stereochemistry, resulting in MAP4C. Both MXFP and MAP4C were integrated into a revised version of the peptide design genetic algorithm (PDGA), a modular, rule-based framework for generating synthetically accessible peptide analogs. Coupling MAP4C to PDGA enabled efficient similarity-based exploration of combinatorial peptide spaces exceeding 10^60 structures. In addition, MXFP could be used to generate pharmacophorically similar peptide analogs of any query structure.

The thesis also explores the use of deep learning models for prediction tasks related to peptides and natural products. A general-purpose language model (GPT-3.5 turbo) was benchmarked against established models for classifying antimicrobial and hemolytic peptide sequences. In a separate project, a transformer-based model was trained to predict the absolute configuration of natural products from achiral molecular input, potentially serving as a computational alternative to experimental stereochemistry assignment.

Table of Contents

1 Thesis Outline

1.1 Thesis Scope and Outline

Most modern drugs available on the market fall within the two main categories of small organic molecules and biologics. In early drug discovery, computational methods for these two groups differ substantially, with small organic molecules typically relying on molecular graph-based representations and biologics typically relying on sequence-based representations. However, emerging therapeutic modalities blur the line between these two categories. Therapeutic peptides, in particular, combine characteristics of both.

For example, many natural antimicrobial peptides (AMPs), such as vancomycin and polymyxin, are primarily composed of peptide monomers but also incorporate a variety of chemical modifications. These modifications are often difficult to encode accurately using conventional sequence-based methods, which are designed for unmodified peptide chains. At the same time, traditional cheminformatics approaches fail to represent the size and sequence-dependent properties of these larger and more complex molecules. As a result, there is a pressing need for computational tools that can not only accommodate the structural scale of therapeutic peptides (and other macromolecules) but also adapt to their chemical richness.

Addressing this need, previous work in the Reymond group led to the development of two molecular fingerprints tailored for macromolecular encoding: the macromolecule extended atom-pair fingerprint (MXFP) and the MinHashed atom-pair fingerprint (MAP4). In addition, a genetic algorithm that uses these fingerprints to guide the generation of peptide analogs from any query structure was introduced. During my PhD, I contributed to both (i) the further development and refinement of these methods and (ii) their integration into practical applications for the design of new AMPs, in close collaboration with the Reymond group's wet lab team. The content of this thesis focuses primarily on objective (i), covering the computational developments in detail. However, the

tools described here have been applied in both completed and ongoing collaborative projects within the research group and with external collaborators (related publications are outlined in section 1.2). The following chapter outline is intended to provide context and a common thread for the individual projects presented in this thesis:

Chapter 2 provides a general introduction to key concepts and techniques relevant to the thesis. It serves as a foundation for understanding the context and contents presented in later chapters.

Chapter 3 began with the goal of introducing an open-source RDKit implementation of MXFP. However, the project evolved into a broader exploration of how tools originally developed for reaction informatics could be repurposed to analyse chemical spaces composed of molecular pairs. This concept is potentially useful for identifying meaningful transformations and scaffold hops within analog series generated by generative models.

Chapter 4 focuses on the development of a chiral extension of the MAP4 fingerprint, resulting in MAP4C. A major challenge in representing natural products and peptides is the presence of multiple stereocenters, where different stereoisomers, such as D- and L-amino acids, can lead to distinct biological properties. MAP4C was designed to address this issue by explicitly encoding stereochemistry, enabling the differentiation of different stereoisomers. The fingerprint retains compatibility with both small and large molecules, making it particularly suitable for the analysis of structurally diverse datasets.

Chapter 5 describes how the newly developed MXFP and MAP4C fingerprints were incorporated into a refined version of the peptide design genetic algorithm (PDGA). The updated algorithm was modularized to support a range of chemically realistic modifications, reflecting both the structural diversity observed in nature and the constraints of synthetic accessibility. The performance of PDGA was evaluated in large combinatorial spaces (on the order of 10^60 structures) demonstrating its

ability to recover arbitrary query compounds and to propose structurally related analogs of these queries.

Chapter 6 focuses on the emerging role of large language models (LLMs) in chemistry. During the initial wave of interest in tools such as ChatGPT, we evaluated whether a general-purpose LLM (GPT-3.5 turbo) could be used out of the box to predict the antimicrobial and hemolytic activity of peptide sequences. Although the model performed comparably to established predictors, its high computational cost limited its practical utility. Nevertheless, the study provided an early benchmark for the potential application of LLMs in molecular activity prediction. Some of the models and approaches developed in this context have also been used to guide candidate selection in both published work and ongoing collaborations within the group.

Chapter 7 continues the exploration of transformer-based models by applying them to a specific challenge in natural product research: the prediction of absolute stereochemistry. A transformer model (NPstereo) was trained to predict the most likely absolute configuration (chirality) of a natural product based solely on its achiral SMILES representation. NPstereo achieved good predictive performance and could potentially offer a computational alternative to experimentally intensive methods for stereochemical assignment.

1.2 Publications

The thesis is based on the first-author publications listed below, each presented as an individual chapter.

1. Orsi, M.; Probst, D.; Schwaller, P.; Reymond, J.-L. Alchemical Analysis of FDA Approved Drugs. *Digital Discovery* **2023**, 10.1039.D3DD00039G. https://doi.org/10.1039/D3DD00039G.

2. Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. *J Cheminf.* **2024**, *16* (1), 53. https://doi.org/10.1186/s13321-024-00849-6.
3. Orsi, M.; Reymond, J. Navigating a 1E+60 Chemical Space of Peptide/Peptoid Oligomers. *Molecular Informatics* **2024**, e202400186. https://doi.org/10.1002/minf.202400186.
4. Orsi, M.; Reymond, J.-L. Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity? *RSC Med. Chem.* **2024**, *15* (6), 2030-2036. https://doi.org/10.1039/D4MD00159A
5. Orsi, M.; Reymond, J.-L. Assigning the Stereochemistry of Natural Products by Machine Learning. *ChemRxiv*, February 7, **2025**. https://doi.org/10.26434/chemrxiv-2024-zz9pw.

The following co-authored publications were produced during the course of this thesis. Relevant contributions have been integrated into the thesis where explicitly indicated.

1. Zakharova, E.; Orsi, M.; Capecchi, A.; Reymond, J. Machine Learning Guided Discovery of Non-Hemolytic Membrane Disruptive Anticancer Peptides. *ChemMedChem* **2022**. https://doi.org/10.1002/cmdc.202200291.
2. Cai, X.; Orsi, M.; Capecchi, A.; Köhler, T.; van Delden, C.; Javor, S.; Reymond, J.-L. An Intrinsically Disordered Antimicrobial Peptide Dendrimer from Stereorandomized Virtual Screening. *Cell Reports Physical Science* **2022**, *3* (12), 101161. https://doi.org/10.1016/j.xcrp.2022.101161.
3. Cai, X.; Capecchi, A.; Olcay, B.; Orsi, M.; Javor, S.; Reymond, J. Exploring the Sequence Space of Antimicrobial Peptide Dendrimers. *Israel Journal of Chemistry* **2023**, *63* (10-11), e202300096. https://doi.org/10.1002/ijch.202300096.
4. Orsi, M.; Shing Loh, B.; Weng, C.; Ang, W. H.; Frei, A. Using Machine Learning to Predict the Antibacterial Activity of Ruthenium Complexes. *Angew Chem Int Ed* **2024**, *63* (10), e202317901. https://doi.org/10.1002/anie.202317901.

5. Frei, A.; Orsi, M. ELECTRUM: An Electron Configuration-Based Universal Metal Fingerprint for Transition Metal Compounds. *ChemRxiv*, October 17, 2024. https://doi.org/10.26434/chemrxiv-2024-vqktn.
6. Orsi, M.; Personne, H.; Bonvin, E.; Paschoud, T.; Olcay, B.; Hu, X.; Javor, S.; Reymond, J.-L. Chemical Space for Peptide-Based Antimicrobials. *Chimia* **2024**, *78* (10), 648-653. https://doi.org/10.2533/chimia.2024.648.
7. Eugster, R.; Orsi, M.; Buttitta, G.; Serafini, N.; Tiboni, M.; Casettari, L.; Reymond, J.-L.; Aleandri, S.; Luciani, P. Leveraging Machine Learning to Streamline the Development of Liposomal Drug Delivery Systems. *Journal of Controlled Release* **2024**, *376*, 1025-1038. https://doi.org/10.1016/j.jconrel.2024.10.065.
8. Bonvin, E.; Orsi, M.; Paschoud, T.; Gopalasingam, A.; Reusser, J.; Köhler, T.; Van Delden, C.; Reymond, J. Antimicrobial Peptide-Peptoid Macrocycles from the Polymyxin B2 Chemical Space. *Angew Chem Int Ed* **2025**, e202501299. https://doi.org/10.1002/anie.202501299.
9. Carrel, A.; Yiannakas, A.; Roukens, J.-J.; Reynoso-Moreno, I.; Orsi, M.; Thakkar, A.; Arus-Pous, J.; Pellegata, D.; Gertsch, J.; Reymond, J.-L. Exploring Simple Drug Scaffolds from the Generated Database Chemical Space Reveals a Chiral Bicyclic Azepane with Potent Neuropharmacology. *J. Med. Chem.* **2025**, acs.jmedchem.4c02549. https://doi.org/10.1021/acs.jmedchem.4c02549.

1.3 Conference Presentations

1. Applied Machine Learning Days, Lausanne, Switzerland, 2022 (oral presentation, poster)
2. Bern Data Science Day, Bern, Switzerland, 2022 (poster)
3. RDKit UGM, Berlin, Germany, 2022 (oral pitch, poster)
4. Bern Data Science Day, Bern, Switzerland, 2023 (poster)
5. RDKit UGM, Mainz, Germany, 2023 (oral pitch, poster)
6. SCS Fall Meeting, Bern, Switzerland, 2023 (poster)
7. Applied Machine Learning Days, Lausanne, Switzerland, 2024 (poster)
8. AI@UniBE, Bern, Switzerland, 2024 (oral pitch)
9. DMCCB Basel Symposium, Basel, Switzerland, 2024 (poster)
10. EuroQSAR, Barcelona, Spain, 2024 (poster)
11. NLP in Chemistry, Bern, Switzerland, 2024 (oral presentation)
12. Novo Nordisk Chemist of the Future, Copenhagen, Denmark, 2024 (oral presentation)
13. Peptide Therapeutics Forum, Basel, Switzerland, 2024 (oral presentation)
14. RDKit UGM, Zürich, Switzerland, 2024 (oral pitch)
15. ACS Spring Meeting 25, San Diego, USA, 2025 (oral presentation)

2 Introduction

2.1 Scaling Drug Discovery: From High-Throughput to *In Silico*

The advent of high-throughput screening (HTS) in the 1990s marked a paradigm shift in drug discovery, transforming it from a largely hypothesis-driven discipline into a numbers-driven endeavour. With the ability to experimentally test hundreds of thousands to millions of compounds against a biological target, HTS established the idea that success could be achieved by simply increasing the volume of screened candidates. Technologies like phage display[1,2] and DNA-encoded libraries (DELs)[3] further pushed the boundaries of this idea, enabling the experimental screening of libraries containing up to 10^9-10^11 variants, which is to date the highest number of compounds screenable *in vitro*.[4-6]

Despite these advancements, *in vitro* screening remains fundamentally limited. Practical constraints such as the cost and time associated with compound synthesis, reagent availability, solubility, and assay throughput restrict the size and diversity of screenable chemical space. More importantly, these methods are unable to access the vast majority of all theoretically plausible molecules, which are collectively referred to as chemical space.[7-11] Although this space is conceptually unlimited, drug discovery has traditionally focused on the subset relevant to drug-like small organic molecules. Even within this constrained region, estimates suggest that chemical space may encompass between 10^20 and 10^60 structures, which is orders of magnitude beyond what any experimental screening library can contain.[12–14]

To address these limitations, theoretical and computational approaches have emerged as powerful alternatives to *in vitro* techniques. Operating entirely *in silico*, they enable the efficient exploration of vast regions of chemical space that are otherwise inaccessible through experimental means. This computational turn has been made possible by the rise of cheminformatics, a discipline that began with the task of digitizing chemical structures and databases, but has since evolved into a comprehensive toolkit for modelling, analysing, and navigating chemical space.[15] Today,

cheminformatics supports nearly every stage of early drug discovery and serves as an interface between experimental data and compound design, enabling the interpretation of experimental results and prioritization of candidates for synthesis and further testing.[16-20]

Central to cheminformatics is the question of how molecular structures can be represented in formats suitable for computational processing. Translating the complexity of chemical structures into machine-readable encodings is the foundation upon which all downstream analyses, such as similarity searching, virtual screening, or predictive modelling, are built.[21–23] The next section introduces how molecules are digitally represented for computational use, and how these representations serve as the basis for cheminformatics and machine learning pipelines.

2.2 Molecular Representations for Computational Applications

To operate on molecular structures, computers require these structures to be translated into a machine-readable format. In cheminformatics, this is typically done by transforming molecules into high-dimensional numerical vectors. These representations, commonly referred to as molecular descriptors or fingerprints, serve as mathematical proxies for molecular identity and summarize structural features such as atom types, connectivity, electronic distribution, stereochemistry, and spatial geometry.[24,25] Importantly, this transformation enables quantitative comparisons between molecules by placing them in a structured, vectorized space where distances and similarities can be defined using formal metrics.[26,27] As such, molecular descriptors form the basis for tasks such as compound clustering, virtual screening, and machine learning applications that require structured numerical input.

Molecular descriptors can be broadly categorized by dimensionality: 1D descriptors typically reflect single physicochemical properties; 2D descriptors account for full molecular topology based on graph representations; and 3D descriptors incorporate spatial geometry, which offers more detailed information but requires 3D coordinates, making them slower to compute and less widely applicable in early-phase screening. For most tasks in drug discovery, 2D descriptors strike the best balance

between detail and computational efficiency.[28] In particular the descriptor classes described below will be relevant for the contents discussed in the thesis (**Figure 1**).

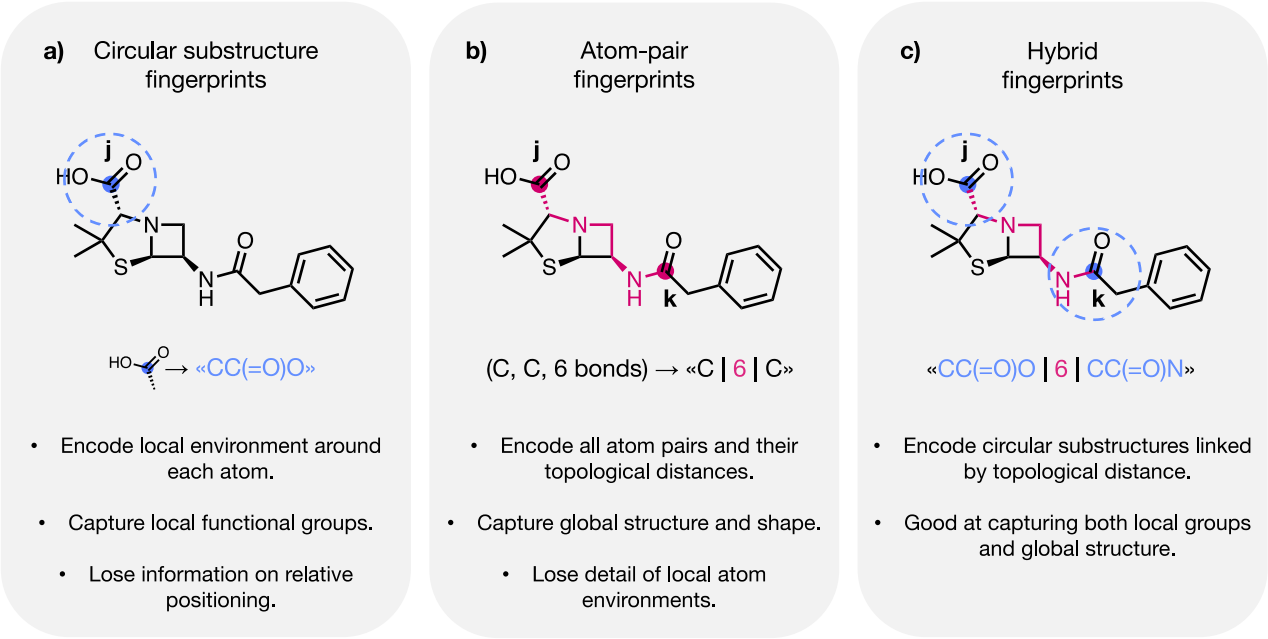


Figure 1. Comparison of common molecular fingerprinting strategies relevant for this thesis. a) Substructure-based fingerprinting encodes atom-centered circular neighborhoods defined by a fixed bond radius. b) Atom-pair fingerprints represent all atom pairs along with their topological distance, capturing global molecular shape. c) Hybrid fingerprints combine circular substructures with atom-pair distances by generating shingles of substructure pairs and their separating topological distance.

2.2.1 Circular substructure fingerprints

Circular substructure fingerprints, such as Extended Connectivity Fingerprints (ECFPs), encode molecular structure by enumerating all atom-centered circular subgraphs within a specified bond diameter (**Figure 1a**).[29,30] Each subgraph is hashed into a numerical identifier, and these identifiers are subsequently folded into a fixed-length binary vector using a modulo operation, yielding a sparse bitstring that serves as a unique structural fingerprint of the molecule. Owing to their lightweight implementation and robust performance, ECFPs are to date the golden standard for similarity-based tasks in cheminformatics.[31]

MinHashed Fingerprints (MHFPs) represent a recent evolution of this concept.[32] Like ECFPs, they extract circular atom environments but encode them as SMILES strings instead of feature-based graphs. These strings are hashed using the SHA-1 algorithm and subjected to MinHashing, a

technique borrowed from natural language processing.[33] A key advantage of MinHashing lies in its compatibility with locality-sensitive hashing (LSH) forests, which enable extremely fast and scalable similarity searches across ultra-large chemical libraries.[34] This makes MHFPs particularly well-suited for applications requiring rapid nearest-neighbour queries in very high data regimes.

While circular substructure fingerprints perform exceptionally well for small molecules, they lose resolution when applied to large or repetitive structures such as peptides or polymers. Their limited radius constrains them to local subgraphs, which can obscure global molecular topology especially in cases with repeating monomers (such as with peptides and glycosides) where not only the presence of certain substructures is important, but also their relative placement within the global structure. To overcome this, alternative fingerprinting strategies have been developed to more effectively capture long-range structural information.

2.2.2 Atom-pair fingerprints

Atom-pair fingerprints provide an effective way to capture the complexity of large molecules such as peptides and other natural products. Unlike circular substructure methods that focus on local neighbourhoods, these descriptors represent molecular structure by recording all pairs of atoms along with the topological distances (measured in bond counts) that separate them (**Figure 1b**). This captures global molecular features, including spatial relationships and overall shape.[35]

The Xfp fingerprint, developed for small molecule comparison, builds on atom-pair representations by additionally classifying atoms according to pharmacophoric features such as hydrogen bond donors (HBD), acceptors (HBA), hydrophobic groups, and planar atoms. It then counts the frequency of all atom-pair combinations across discrete topological distances, typically ranging from one to ten bonds.[36,37] Building on this concept, the Macromolecule Extended Atom-Pair Fingerprint (MXFP) adapts this encoding for larger, more complex structures by expanding both the distance resolution (using 31 bins that cover a wide range of bond lengths) and the pharmacophoric groups (including six distinct functional categories beyond simple atom identity).[38,39] This encoding

enables the fingerprint to retain long-range information and capture key functional patterns found in macromolecules while still maintaining a manageable representation size and computation time.

2.2.3 Hybrid Fingerprints

While circular substructure and atom-pair fingerprints each offer unique advantages, they also come with trade-offs: the former excels at capturing local environments but struggles with large-scale topology and sequence-dependent effects, while the latter encodes long-range features at the cost of lower local precision. Hybrid fingerprints aim to unify these approaches by combining their respective strengths (**Figure 1c**).

One such fingerprint is the MinHashed Atom-Pair fingerprint up to four bonds (MAP4), which combines the methods of circular substructure and atom-pair fingerprints.[40] MAP4 extracts atom-centered circular substructures up to two bonds away and converts these into canonical SMILES strings. From this set of substructures, MAP4 generates all possible atom pairs and encodes them as shingles, composite strings that combine the two substructure SMILES with their topological distance in the format: "SMILES1|dist|SMILES2". These shingles are then MinHashed into a fixed-length fingerprint. A key strength of MAP4 lies in its versatility: it performs well across a wide range of molecule sizes, from small drug-like compounds to large peptides and macrocycles. The MinHashing step ensures compatibility with LSH, allowing rapid nearest-neighbour searches and scalable chemical space visualization using algorithms like TMAP.[41]

To further expand its applicability, the MAP4C variant extends the MAP4 encoding to incorporate chiral information by embedding Cahn-Ingold-Prelog (CIP) descriptors into the shingles prior to hashing.[42] Stereochemistry plays a critical role in molecular function across both small molecules and larger macromolecules: while enantiomers of small drugs can exhibit markedly different pharmacokinetics or target affinities, stereochemical variation in peptides and natural products can impact folding, membrane interaction, and bioactivity.[43,44] To address this, we developed MAP4C as a practical solution for our own experimental workflows, enabling the differentiation of peptide diastereomers containing both D- and L-residues and ensuring that

stereochemical differences are preserved in similarity comparisons. At the same time, MAP4C remains broadly applicable to a wide range of molecular structures, from small chiral drugs to large, stereochemically complex natural products.

2.3 From Representation to Prediction: Machine Learning in Molecular Design

Once molecular structures are encoded numerically, they can be positioned in a structured space where molecular relationships become quantifiable. This enables not only similarity-based operations, but also the application of statistical models that learn from data and generalize to unseen compounds. Machine learning (ML) methods provide a practical solution to this need, offering a way to relate molecular features to experimental outcomes. As datasets have expanded and become more heterogeneous, ML has proven useful for identifying patterns within these datasets that are not easily captured by rule-based heuristics alone.

2.3.1 Descriptor-Based Modelling with Classical Algorithms

One of the earliest and most widely used applications of machine learning in drug discovery involves pairing molecular descriptors with classical algorithms to predict biological activity, toxicity, or target binding.[44-47] Among the first algorithms adopted in this setting were Support Vector Machines (SVMs), Random Forests (RFs), and Multilayer Perceptrons (MLPs). These models, while relatively simple by today's standards, remain effective when combined with well-curated molecular descriptors. SVMs operate by identifying an optimal hyperplane that separates classes in a high-dimensional feature space.[49] RFs are ensemble methods that aggregate the predictions of multiple decision trees, offering robustness to noise and built-in feature importance metrics.[50] MLPs, as simple feedforward neural networks, are capable of learning non-linear mappings from input features to target properties, though they tend to require more tuning and regularization than SVMs or RFs.[51]

All three model types benefit from the fixed length, vectorized nature of molecular descriptors, which allows them to be integrated directly with minimal preprocessing. Their relatively low

computational demands and moderate data requirements make them well suited for applications where experimental throughput is limited. Even in the context of more complex architectures, these models continue to serve as useful baselines and are often preferred in cases where interpretability, fast iteration, or robustness under data constraints are essential.[52,53]

2.3.2 Learning Representations from Molecular Structure

While molecular fingerprints have enabled efficient modelling across many tasks, their design is inherently static. They capture features based on predefined rules and heuristics, which makes them well-suited for similarity comparisons and classical modelling approaches. However, because these features are selected in advance, they may not reflect the structural patterns most relevant to a specific task or dataset. Representation learning offers a more flexible alternative. Instead of relying on hand-crafted descriptors, models can be trained to learn internal molecular representations directly from structural input. These learned representations, or embeddings, are continuous vectors that summarize the chemical and structural information most predictive of a given outcome.[53-55] They are optimized during training and can later be reused for related tasks such as clustering, similarity analysis, or chemical space visualization.

A natural choice for learning such representations is to work directly with the molecular graph, where atoms are treated as nodes and bonds as edges. This format preserves the structure of molecules and avoids the abstraction steps required by traditional fingerprints. Graph neural networks (GNNs) are designed to operate on this kind of data.[56-61] They use iterative message-passing to update atom-level features based on the local connectivity within the graph. Over multiple layers, the model incorporates information from larger regions of the molecule, allowing it to capture both local substructures and broader topological features. Because GNNs learn representations directly from molecular structure, they eliminate the need for manual feature design and can adapt to the statistical properties of the training data. However, graph-based models also come with limitations. They can be computationally intensive, particularly for large molecules, and often require careful tuning of

architecture and message-passing schemes to perform well. In addition, they lack the straightforward parallelization and scalability that have made sequence-based models highly effective.

2.3.3 Sequence-Based Representations and Transformer Models

An alternative is to represent molecules as linear sequences, such as SMILES strings or amino acid chains. These formats make it possible to apply powerful and computationally efficient techniques originally developed for natural language processing (NLP), where models learn to recognize patterns in sequential data.

Early NLP applications in cheminformatics were driven by Recurrent Neural Networks (RNNs), which process input sequences one token at a time. Because chemical structures can be represented as SMILES strings, and peptides as residue sequences, RNNs can be applied with little preprocessing to these string representations directly. RNNs are useful for both predictive and generative tasks.[62-66] However, they suffer from several limitations. Their strictly sequential processing hinders parallelization and makes training inefficient, particularly for long sequences. In addition, their ability to capture dependencies between distant tokens is often limited by vanishing gradients and restricted memory capacity, which can affect performance on inputs with long-range interactions.

The transformer architecture was introduced to overcome the limitations of RNNs, most notably in the seminal paper "Attention Is All You Need", which proposed a fully attention-based model for sequence learning.[68] Unlike RNNs, which process sequences in a strictly sequential manner, transformers apply a self-attention mechanism[69] that allows the model to consider all positions in a sequence simultaneously. This architecture makes it possible to model dependencies between distant tokens without being constrained by input order. As a result, transformers are particularly effective at capturing long-range relationships within molecular sequences, such as correlations between distant functional groups or the influence of backbone modifications in peptides.

When applied to SMILES strings or amino acid sequences, transformers treat molecules as sequences of discrete tokens (typically atoms, bonds, or residues). The self-attention mechanism

enables the model to learn which parts of the sequence are most informative for a given task. Transformers have proven particularly effective in generative modelling.[69-71] By learning the conditional probability of each token given the preceding context, they can be trained to produce new molecular or peptide sequences that follow the syntactic rules of the input representation and (potentially) reflect learned structure-function relationships. When properly trained, these models are capable of generating chemically valid SMILES strings or biologically plausible sequences, conditioned on specific design objectives such as activity, toxicity, or target selectivity.[72-75]

2.3.4 Synergies Between Classical and Deep Learning Approaches

Despite the increasing sophistication of deep learning methods, classical machine learning remains a relevant and complementary approach in many areas of drug discovery. It continues to offer practical advantages in settings where data is limited, or interpretability is a priority. These conditions are common in exploratory or early-phase projects.[17,76-79]

A key distinction between classical machine learning and modern deep learning approaches lies in how they handle molecular input and data availability. Classical models based on molecular fingerprints typically operate on fixed-length, chemically meaningful descriptors. These features encode established structural principles and often act as inductive biases, allowing the models to perform reliably even on relatively small, well-curated datasets and in out-of-distribution scenarios. In contrast, deep learning models usually operate on raw or minimally processed molecular representations (such as graphs or sequences) and require larger, more diverse datasets to learn useful abstractions. While deep models can capture complex, nonlinear relationships between structure and activity, their performance often depends on how well the training data cover the relevant chemical space. When this space is narrow or unrepresentative, generalization becomes more difficult and model reliability can suffer.[80-82] Transfer learning offers a partial solution to this problem by allowing pretrained models to be fine-tuned on smaller, domain-specific datasets.[83-86] However, this strategy still depends on the assumption that the source data contain at least some relevant structural patterns.

Hence, the choice between classical descriptors and learned representations is best viewed in terms of alignment with the problem setting. For tasks with well-defined objectives and limited data, classical machine learning methods often perform competitively, with lower computational requirements, faster deployment, and greater interpretability.[77-79] Deep learning models, by contrast, are well suited to settings where large, diverse datasets are available and the goal is to model complex structure-activity relationships, integrate multiple prediction tasks, or generate novel compounds from learned patterns.[62-66,69-71] Taken together, these considerations highlight that the choice between classical and deep learning approaches should not be based on perceived methodological superiority, but on how well each strategy fits the specific constraints and goals of a given project. In drug discovery and development, where modelling is only one part of a broader experimental (and regulatory) pipeline, factors such as data availability, interpretability, and compatibility with downstream requirements often play a decisive role. Hence, placing the right tool in the right setting is particularly important, since the aim is not just to improve predictions, but to support the discovery of compounds that can ultimately advance toward becoming safe and effective drugs (**Figure 2**).[88,89]

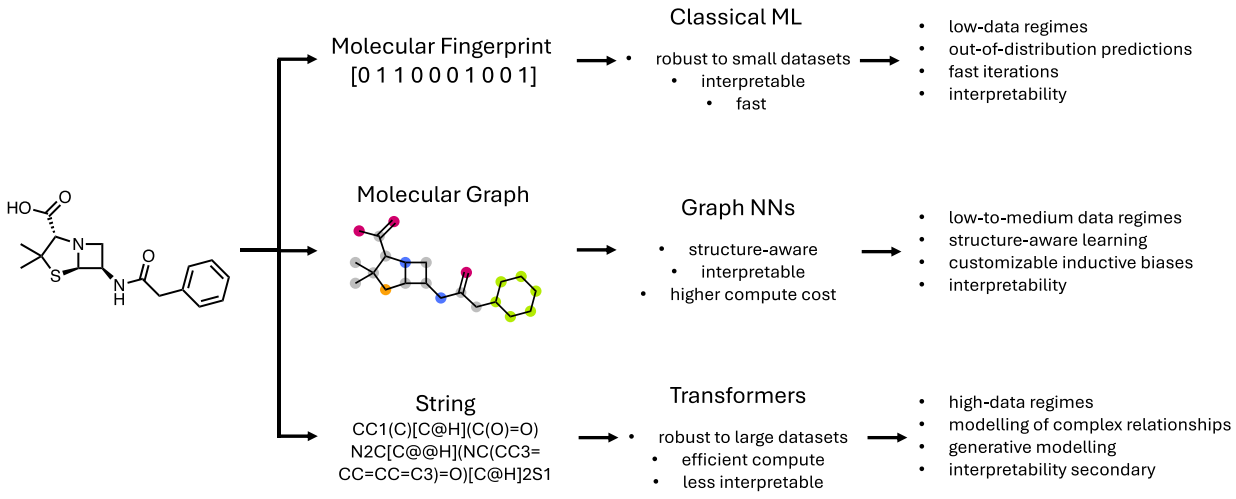


Figure 2. Overview of molecular representation strategies and corresponding machine learning models. The choice of model should reflect both the data regime and the goals of the application.

This framing is particularly relevant in therapeutic areas where conventional discovery strategies have become less effective, and in resource-limited settings such as academic research, where methodological efficiency is essential. Antimicrobial drug discovery exemplifies both

challenges: it demands new molecular strategies while operating under limited funding, low commercial incentive, and the need for targeted, data-efficient design.

2.4 Antimicrobial Resistance and Antimicrobial Peptides

The introduction of antibiotics in the early 20th century transformed the landscape of modern medicine. These compounds, largely discovered from microbial natural products, enabled the treatment of previously fatal bacterial infections and underpinned the safe development of surgical procedures, cancer therapies, and organ transplantation. However, their widespread and often indiscriminate use has led to the emergence of antimicrobial resistance (AMR), a phenomenon in which bacteria evolve mechanisms to survive exposure to antibiotics.[89-91] Over time, resistance has accumulated not only within individual strains but also across species through horizontal gene transfer, leading to a growing number of multidrug-resistant (MDR) pathogens.

Particularly concerning are pathogens in the ESKAPE group (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp*.), which are responsible for a disproportionate share of hospital-acquired infections and are increasingly resistant to last-line therapies.[93] Meanwhile, the development of new antibiotics has slowed dramatically, in part due to scientific challenges in discovering novel chemical scaffolds, and in larger part due to economic disincentives: antibiotics are typically used for short durations and are often conserved to delay resistance, resulting in poor return on investment compared to chronic disease therapeutics. As a result, the pipeline for novel antimicrobial agents remains sparse.[94]

To address this gap, attention has increasingly turned toward non-traditional antimicrobial modalities, among which antimicrobial peptides (AMPs) represent a particularly promising class.[95] AMPs are short, often cationic peptides that are produced by a wide range of organisms, either as components of the innate immune system[96] or as competitive factors in microbial ecosystems.[97] Unlike classical antibiotics that typically target specific enzymes or biosynthetic pathways, AMPs tend to act through more generalized mechanisms, including membrane disruption and aggregation of

intracellular targets.[98,99] These features make them inherently less susceptible to some resistance mechanisms. Despite these advantages, AMPs face a distinct set of challenges that have limited their clinical adoption. Chief among these are:

- Proteolytic instability, leading to rapid degradation by host or bacterial proteases.
- Hemolytic and cytotoxic effects, particularly at concentrations close to the therapeutic window.
- Poor pharmacokinetics, including short half-life and low bioavailability.
- Synthetic complexity, especially for cyclic or modified backbones.

Overcoming these barriers requires systematic exploration of peptide chemical space. This space grows combinatorially with peptide length and diversity, quickly exceeding what can be addressed through traditional synthesis and experimental screening alone.[100] In addition, the landscape of synthetically accessible peptides is constantly evolving, driven by new building blocks and chemistries emerging from experimental practice. As a result, efficient computational methods are needed not only to search vast design spaces, but also to adapt quickly to new synthetic possibilities and guide experimental efforts toward tractable and promising candidates.

One potential strategy for exploring peptide space is the use of generative models, such as those based on transformer-based architectures. Methods like PepINVENT, for example, have shown how models trained on virtual peptide libraries can be used to conditionally generate novel sequences.[101] However, the structural rules governing peptide assembly are relatively simple and synthetically tractable peptides can be validly constructed by well-defined bond-forming rules such as peptide bond formation, disulfide bridging, head-to-tail cyclization, or side-chain crosslinking. As a result, there is less need for generative deep learning models to ensure syntactic validity, unlike in small-molecule design where atom type and connectivity constraints for “drug-likeness” are non-trivial.

Hence, among the tools developed to support peptide design, rule-based frameworks remain particularly attractive due to their adaptability. In the Reymond group, we have introduced the Peptide Design Genetic Algorithm (PDGA), a modular generator designed to explore synthetically feasible peptide spaces using predefined connection rules and customizable building blocks (**Figure 3**).[100,102] Because it relies on explicit chemical rules rather than learned syntax, the PDGA can accommodate new monomers or synthesis strategies without requiring model retraining. Molecular fingerprints, such as MAP4C[42] and MXFP,[38,39] are then used in combination with the PDGA to guide compound generation and similarity-based selection. This approach has been applied in a real-life setting to design polymyxin-inspired peptide-peptoid hybrids, several of which were synthesized and shown to retain antimicrobial activity.[103]

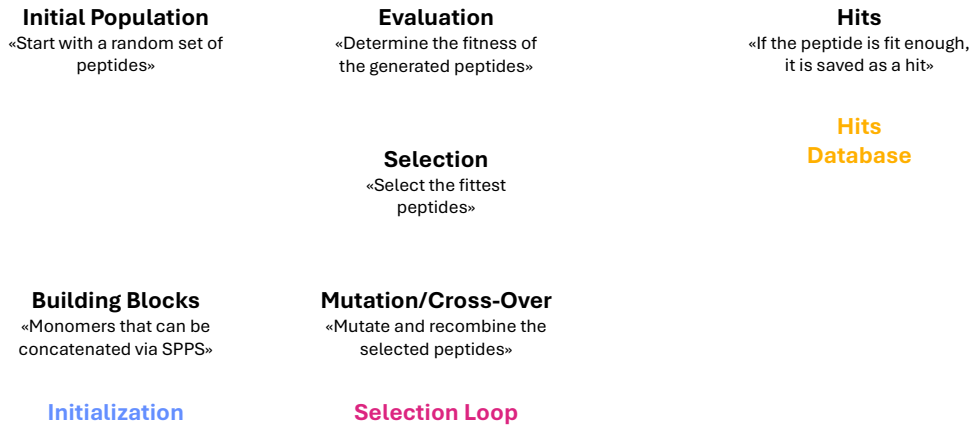


Figure 3. Conceptual overview of the Peptide Design Genetic Algorithm (PDGA). The algorithm is initialized with a fixed set of building blocks and synthetic rules to generate candidate structures. Within the selection loop, candidates are iteratively refined based on a fitness function. High-scoring compounds are retained according to a predefined threshold and stored in a hit database.

Yet many of the challenges associated with AMP design, such as hemolytic activity, proteolytic stability, or serum half-life, cannot be anticipated from structural similarity alone. For these properties, predictive models trained on experimental data are essential.[65,66,103-107] Integrated into generative workflows, such models can filter or re-rank sequences based on complex biological properties. As synthesis platforms become more automated and peptide datasets continue to grow, the ability to iteratively generate, evaluate, and refine candidate structures will become increasingly central to antimicrobial discovery. Within this context, combining rule-based design with predictive

modelling provides a practical and scalable foundation. The next chapters of this thesis will focus on the development, implementation, and evaluation of computational methods that support this goal.

3 Alchemical analysis of FDA approved drugs

This chapter is based on a scientific article previously published in *Digital Discovery*. The article is reproduced here under the terms of the Creative Commons Attribution License (CC BY 3.0):

Orsi, M.; Probst, D.; Schwaller, P.; Reymond, J.-L. Alchemical Analysis of FDA Approved Drugs. *Digital Discovery* **2023**, 10.1039.D3DD00039G. https://doi.org/10.1039/D3DD00039G.

Abstract

Chemical space maps help visualize similarities within molecular sets. However, there are many different molecular similarity measures resulting in a confusing number of possible comparisons. To overcome this limitation, we exploit the fact that tools designed for reaction informatics also work for alchemical processes that do not obey Lavoisier's principle, such as the transmutation of lead into gold. We start by using the differential reaction fingerprint (DRFP) to create tree-maps (TMAPs) representing the chemical space of pairs of drugs selected as being similar according to various molecular fingerprints. We then use the Transformer-based RXNMapper model to understand structural relationships between drugs, and its confidence score to distinguish between pairs related by chemically feasible transformations and pairs related by alchemical transmutations. This analysis reveals a diversity of structural similarity relationships that are otherwise difficult to analyse simultaneously. We exemplify this approach by visualizing FDA-approved drugs, EGFR inhibitors, and polymyxin B analogs.

3.1 Introduction

Mapping molecular databases in a chemical space where distances represent similarities between molecules helps to understand their structural similarities and identify relationships that can provide critical insights for drug development and related fields.[108-122] However, molecular similarity can be computed in multiple ways,[124,125] typically using various molecular fingerprints,[22] resulting in a confusing multiplicity of possible chemical space representations.[37,126]

To overcome this limitation and create a chemical space map considering various similarity measures simultaneously, we report a new approach of applying reaction informatics tools to map and analyze drug pairs, namely the differential reaction fingerprint (DRFP)[127] and the Transformer-based RXNMapper model,[127-129] respectively **(Figure 4)**. These tools were initially designed to analyze chemical reactions. However, they can also be applied to processes that do not obey Lavoisier's principle, the conservation of mass, such as the alchemical transmutation of lead into gold.[131,132] Here, we apply them to transmutations between pairs of molecules selected for their similarity according to various molecular fingerprints as similarity measures, an approach related to the recent development of transformer models for drug optimization.[133,134]

We start by using DRFP, which encodes chemical reactions by storing the symmetric difference of two sets containing the circular molecular n-grams generated from the molecules of the molecular pair as a binary fingerprint,[127] to represent the chemical space of drug pairs as a TMAP (tree-map).[41] A TMAP lays out the minimum spanning tree of the nearest neighbour graphs according to a selected similarity measure, here DRFP, and represents a remarkably efficient dimensionality reduction method for high-dimensional datasets. The DRFP TMAP visualization provides a global similarity perspective across drug pairs combining the selected similarity measures. We then use RXNMapper,[128] a model trained on one million reactions documented in the USPTO dataset[135] to pair corresponding atoms between reactants and products in a chemical reaction, to identify the structural relationship between drugs. The confidence score of this transformer appears not to correlate with any of the molecular similarity measures used. It allows us to distinguish drug pairs

related by feasible chemical processes, such as matched molecular pairs corresponding to substituent exchanges,[136,137] from those related by more esoteric, alchemical transmutations including scaffold-hopping changes.[138,139] We demonstrate this approach with the example of FDA-approved drugs as a diversity set, as well as for a series of EGFR inhibitors and polymyxin B analogs as two high similarity sets chosen among small molecule drugs and peptide macrocyclic drugs, respectively.

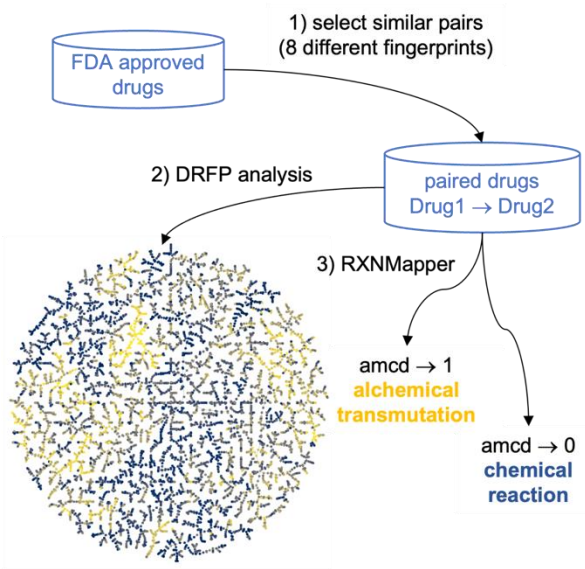


Figure 4. Principle of alchemical analysis of molecular sets at the example of FDA approved drugs. 1) Drugs pairs passing a similarity threshold according to eight different molecular fingerprints are selected. 2) The set of selected pairs is mapped in a TMAP computed using the differential reaction fingerprint (DRFP), color coded by the RXNmapper confidence distance (amcd). 3) the amcd distinguishes pairs of drugs related by a possible reaction (amcd → 0) from those related by an alchemical transmutation (amcd → 1).

3.2 Methods

3.2.1 Datasets

The set of FDA-approved drugs was downloaded from ZINC15,[140,141] the SMILES were canonicalized and kekulized and duplicates were removed to obtain a set of 1,213 unique chemical structures. For the EGFR set, all compounds binding to the tyrosine kinase erbB1 with a molecular weight <700 and an annotated IC_{50} value were downloaded from ChEMBL-31.[142] After SMILES canonicalization and kekulization, duplicates were removed and the 1,500 molecules with the highest ECFP4 Tanimoto similarity to afatinib were selected for the final set. The polymyxin B similarity set

was downloaded from ChEMBL-31 by selecting compounds above the 55% ChEMBL similarity threshold with annotated MIC values. The SMILES were canonicalized and kekulized, and duplicates were removed resulting in a final set of 274 structures.

3.2.2 Molecular fingerprints and similarity calculations

Chemical structures were encoded as eight different fingerprints, namely extended connectivity fingerprints ECFP4 and ECFP6,[29,30] the MinHashed Fingerprint MHFP6,[32] the RDKit Atom-Pair Fingerprint (AP),[35] the Macromolecule Extended Fingerprint (MXFP),[38] the MinHashed Atom-Pair fingerprint MAP4,[40] the Molecular ACCess System keys (MACCS),[143] and Molecular Quantum Numbers (MQNs).[144] ECFP4, ECFP6, AP, MACCS and MQN were calculated using the implementation in the RDKit package (2022.3.4. , https://www.rdkit.org). ECFPs were calculated as 2048-bit vectors. MHFP6 and MAP4 were calculated as 2048-bit vectors using the code described in https://github.com/reymond-group/mhfp and https://github.com/reymond-group/map4. MXFP was calculated using a new open-source version available at https://github.com/reymond-group/mxfp_python. The differential reaction fingerprint (DRFP)[127] was calculated as 2048-bit vectors using the code available at https://github.com/reymond-group/drfp.

Pairwise distances for every possible molecular pair were calculated and stored as a matrix for each fingerprint. Distances were calculated as Jaccard distances (d_J) for ECFP4, ECFP6, MHFP6, AP, MAP4 and MACCS keys, and as Taxicab distances (d_T) for MXFP and MQNs, with values min-max standardized. We selected similar pairs by applying the following distance threshold: $d_J < 0.6$ for ECFP4, ECFP6, MHFP6, $d_J < 0.5$ for AP, $d_J < 0.2$ for MACCS, $d_J < 0.8$ for MAP4, $d_T < 0.1$ for MXFP and $d_T < 0.05$ for MQN (Taxicab distances after rescaling) for the FDA set and $d_J < 0.2$ for ECFP4, ECFP6, MHFP6, AP, $d_J < 0.0125$ for MACCS, $d_J < 0.3$ for MAP4, $d_T < 0.1$ for MXFP and $d_T < 0.05$ for MQN for the EGFR and PMB sets.

Additionally, the ranking of molecular pairs for every compound and fingerprint was calculated, resulting in 1,213 ranked lists of 1,213 pairs each for the FDA set, 1,500 ranked lists of 1,500 ranked

pairs for the EGFR set and 274 ranked lists of 274 pairs for the polymyxin B similarity set for each fingerprint.

Violin plots to display the distribution of distances for every fingerprint and heatmaps to visualize correlations between fingerprints were generated using the seaborn (0.11.2) package. The pairwise distance distributions were balanced out by calculating the ranking of molecular pairs for every compound, resulting in 1,213 ranked lists of 1,213 pairs each for the FDA set, 1,500 ranked lists of 1,500 ranked pairs for the EGFR set and 274 ranked lists of 274 pairs for the polymyxin B similarity set.

3.2.3 Reaction informatics

A reaction SMILES in the form "SMILES1>>SMILES2" (forward reaction) as well as "SMILES2>>SMILES1" (backward reaction) was generated for every selected molecular pair. The forward reaction SMILES was generated to always have the molecule with the lower heavy atom count as a reactant and the molecule with the higher heavy atom count as a product. The reaction SMILES for each drug pair was then encoded using DRFP.[127] The 20 nearest neighbors (NNs) in the DRFP feature space were extracted and the minimum spanning tree layout calculated using the TMAP package.[41] The resulting layout was displayed interactively using Faerun.[145] In addition, the atom-mapping and the corresponding atom-mapping confidence scores were computed for each drug pair reaction SMILES using the published model described in the RXNmapper[128] GitHub repository https://github.com/rxn4chemistry/rxnmapper.

3.3 Results and Discussion

3.3.1 Datasets and selection of drug pairs

To test our reaction informatics approach to map drug space, we selected 1,213 FDA-approved drugs as a representative high diversity set. As examples of a more focused series, we accessed the ChEMBL database[142] and retrieved 1,500 analogs of the small molecule drug afatinib, a kinase inhibitor blocking the endothelial growth factor receptor (EGFR) and used to treat non-small cell lung

carcinoma (NSCLC),[146] as well as 274 analogs of polymyxin B (PMB), an FDA-approved macrocyclic peptide natural product considered as a last resort antibiotic against multidrug-resistant bacteria.[147]

To represent molecular similarities, we considered three types of molecular fingerprints. First, we selected the classical Morgan fingerprint,[30] also called extended connectivity fingerprint (ECFP),[29] which is a binary fingerprint encoding the presence of specific atom-centered circular substructures up to a diameter of four (ECFP4) and six (ECFP6) bonds, as well as our recently reported MinHashed fingerprint MHFP6,[32] which similarly encodes circular substructures up to a diameter of six bonds using shingling and MinHashing to compress information.[148] These circular substructure fingerprints are particularly efficient in virtual screening benchmarks[24,32] and off-target prediction tasks.[149,150] Second, we considered three pharmacophore fingerprints encoding the relative positions of atoms in a molecule and representing molecular shape, namely the RDKit atom-pair fingerprint AP,[35] our recently reported macromolecule extended atom-pair fingerprint MXFP,[38] and the MinHashed Atom-pair fingerprint up to a diameter of four bonds MAP4.[40] Finally, we also included two composition fingerprints, namely MACCS keys[143] and molecular quantum numbers (MQN),[144] which encode the presence and number of features present in a molecule.

To identify relevant pairs in each of our three drug sets (FDA, EGFR and PMB), we computed all pairwise distances in each fingerprint as either Jaccard distance d_J (ECFP4, ECFP6, MHFP6, AP, MAP4, MACCS keys) or Taxicab distance d_T (MXFP, MQN). For all fingerprints, distance zero indicates highest similarity. For each molecule in each set, we then selected the NN for each of the eight fingerprints, as well as any molecule appearing in at least seven of the eight lists of top-20 nearest neighbors. In addition, we selected all drug pairs having a certain similarity in each fingerprint by applying a maximum Jaccard distance (d_J) threshold (see Methods for details).

This selection corresponded to 6,406 (0.87 %) of the 735,078 possible drug pairs in the FDA set, 8,932 (0.79 %) of the 1,124,250 possible drug pairs in the EGFR set, and 8,464 (22.63 %) of the 37,401 possible drug pairs in the PMB set. Each drug was represented in the selected pairs between 1 and 193 times in the FDA approved set, between 1 and 870 times in the EGFR set, and between 4

and 1,031 times in the PMB set (**Figure A1**). Compared to the exhaustive list of drug pairs, the selected drug pairs were enriched in high similarity pairs with lower values of Jaccard distance (d_J). They spanned the entire similarity range in each fingerprint, reflecting the fact that the different fingerprints captured different similarity features (**Figure 5a/Figure 6a/Figure 7a**). Distances were correlated between ECFP4, ECFP6, MHFP6, MAP4, which all encode circular substructures around atoms ($r^2 \sim 0.8$, **Figure 5b/Figure 6b/Figure 7b**). The correlations of MAP4 with other circular substructure fingerprints, particularly in the polymyxin B2 set, were generally lower. This can be attributed to its hybrid nature, which encodes both substructures and atom-pairs. Even so, the correlation between MAP4 and circular substructure fingerprints is notably stronger than its correlation with other fingerprint types. AP and MACCS, which both encode atomic features, were weakly correlated with each other and to a lesser extent with circular fingerprints ($r^2 \sim 0.5$). Finally, MQN and MXFP distances were partly correlated with each other ($r^2 \sim 0.5$) but not with any other fingerprints, probably because both fingerprints are size-dependent and count similar features in molecules.

3.3.2 DRFP chemical space maps

To gain a closer insight into the pairwise relationships among the selected drug pairs, we represented each pair in the form of a reaction SMILES considering the conversion of one drug into the other. Form the reaction SMILES, we then computed the differential reaction fingerprint (DRFP),[127] which encodes the circular substructures that occur only in either the reactant or the product. To represent the DRFP chemical space illustrating the similarities between different drug pairs, we then computed a tree-map (TMAP) providing an overview of drug pairs in each of the three datasets, using various color codes to visualize pair properties (**Figure 5c/Figure 6c/Figure 7c**). The TMAP of DRFP similarities organized pairs by structural types, often series of close analogs of a reference drug. Furthermore, in the FDA-approved drug set, different compound families such as amino acids, steroids, β-lactams, catecholamines, benzodiazepines or prostaglandins appeared in different regions

of the map. This was visible upon close inspection of the interactive TMAPs and is illustrated here for the FDA drug set with the color $FCsp^3$ (**Figure 5c**).

Interactive browsing of the TMAPs made it very easy to inspect drug pairs with specific properties. For example, with the EGFR set, color-coding by activity differences pointed to the few similar drug pairs representing activity cliffs (**Figure 6c**). Inspection of TMAPs was also key to identifying interesting pairs from the point of view of their transformations, as discussed below.

3.3.3 Atom mapping

To estimate whether paired drugs were interconvertible by a feasible chemical reaction or required a more esoteric transmutation, we subjected the drug pair reaction SMILES to the Transformer-based RXNMapper model,[128] which returns an atom-to-atom comparison illustrating the structural relationships within pairs, as well as an atom-mapping confidence score. Atom-mapping confidence scores were determined for the forward and backward reactions and converted to atom-mapping confidence distances (amcd), defined here as one minus the confidence score. In most cases the amcd values were similar for forward and backward reactions, however since the difference was sometimes substantial (**Figure A2**), we used the mean amcd of forward and backward reactions for our analysis. The mean amcd value spanned the entire range between low and high distance (last entry, **Figure 5a**/**Figure 6a**/**Figure 7a**) except for the PMB set, which mainly contains high confidence distances as the structures are too big for the model to map with high confidence. Further, the amcd was not correlated with any of the selected molecular similarities (last entry, **Figure 5b**/**Figure 6b**/**Figure 7b**).

Low amcd values indicated drug pairs related by a simple and usually feasible chemical transformation, usually a functional group change or addition as those found in matched molecular pairs,[136,137] illustrated in the FDA set for the hydroxylation of L-tyrosine to L-DOPA (**Figure 5d**), and in the EGFR set for a Suzuki coupling resulting in a large activity change (**Figure 6d**). In the case of the PMB set, low amcd values indicated pairs related by single amino acid exchange often potentially corresponding to a reaction, for example mutation of a glycine to a phenylalanine residue

corresponding formally to an α-alkylation of glycine with benzyl bromide **(Figure A6**). This observation suggests that the amcd metric effectively captures chemically intuitive transformations, aligning well with the way chemists predict and perceive such changes in molecules during drug design and development.

On the other hand, high amcd values indicated alchemical transmutations that cannot be realized easily, such as scaffold-hopping changes.[138,139] Note that the RXNMapper assigned corresponding atoms mostly in a correct manner even for pairs giving high amcd values. For example, tetrabenazine is paired with hydrocodone by seven of the eight molecule fingerprints used for pairing. The transformation features an exotic double-ring formation accompanied by a reshuffling of the 23 atoms (**Figure 5e**). A similarly exotic alchemical change relates afatinib with osimertinib, an analog matched by all eight fingerprints used for pairing (**Figure 5f**). In the EGFR set, a double linker modification preserving activity relates CHEMBL469997 to CHEMBL181275, whereby the benzyl ether linker is obtained by combining an oxygen atom of the sulfone with a methylene group of the aminobutanol second linker group (**Figure 6e**). In another scaffold hopping change between CHEMBL469997 and CHEMBL181275, an aniline substituent is incorporated into the adjacent bicyclic system to form a condensed tricyclic heteroaromatic group, resulting in an interesting activity increase (**Figure 6f**).

In the case of the PMB set, many pairs were generally related by high amcd values, probably because the changes corresponded to multiple amino acid exchanges, which cannot be realized on the complete molecules since each sequence analog requires a separate synthesis. Interestingly, one of the high amcd changes corresponds to a simple exchange of four aromatic aldehyde imines attached to the four diaminobutanoic acid residues, a reaction which would seem to be feasible (**Figure 7d**). This imine exchange is however accompanied by a mutation of a leucine residue to a phenylalanine.

Taken together, the analysis of the TMAP of similar drug pairs guided by DRFP similarity and amcd values allowed a rapid insight into multiple interesting comparisons between molecules in each of the three sets analyzed. Further examples of interesting pairs in the FDA approved set are provided in the Supporting information (**Figure A7**).

3.4 Conclusion

In summary, we have shown that borrowing tools from reaction informatics provides an opportunity to map multiple similarity relationships between molecules simultaneously and gain insights into interesting drug pairs that are otherwise difficult to identify. Specifically, we used DRFP to map the chemical space of multiple drug pairs selected as being similar according to eight different molecular fingerprints simultaneously in the form of TMAPs. We then used RXNMapper to visualize the structural changes between drugs and identify pairs of drugs related by feasible chemical transformation from pairs related by alchemical changes corresponding to multiple and complex structural rearrangements. These tools should generally be applicable to analyze drug sets from multiple angles in the context of drug discovery. For instance, they present a promising opportunity to visualize chemically feasible transformations, aiding in the improvement of potential drug candidates. Furthermore, this method offers a potential avenue to detect molecular transformations that enhance biological activity, assisting in determining their chemical feasibility. Lastly, these tools can be applied in prodrug development, facilitating the visualization and identification of potential transformations that a prodrug may undergo, and thus guiding the selection of optimal groups for prodrug design.

3.5 Code availability

The source codes and datasets used for this study are available at https://github.com/reymond-group/alchemical_pairs.

3.6 Author contributions

MO designed and realized the project and wrote the paper. DP provided support for the DRFP implementation and wrote the paper. PS provided support for the RXNmapper implementation and wrote the paper. JLR designed and supervised the project and wrote the paper. All authors read and approved the final manuscript.

3.7 Acknowledgements

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076).

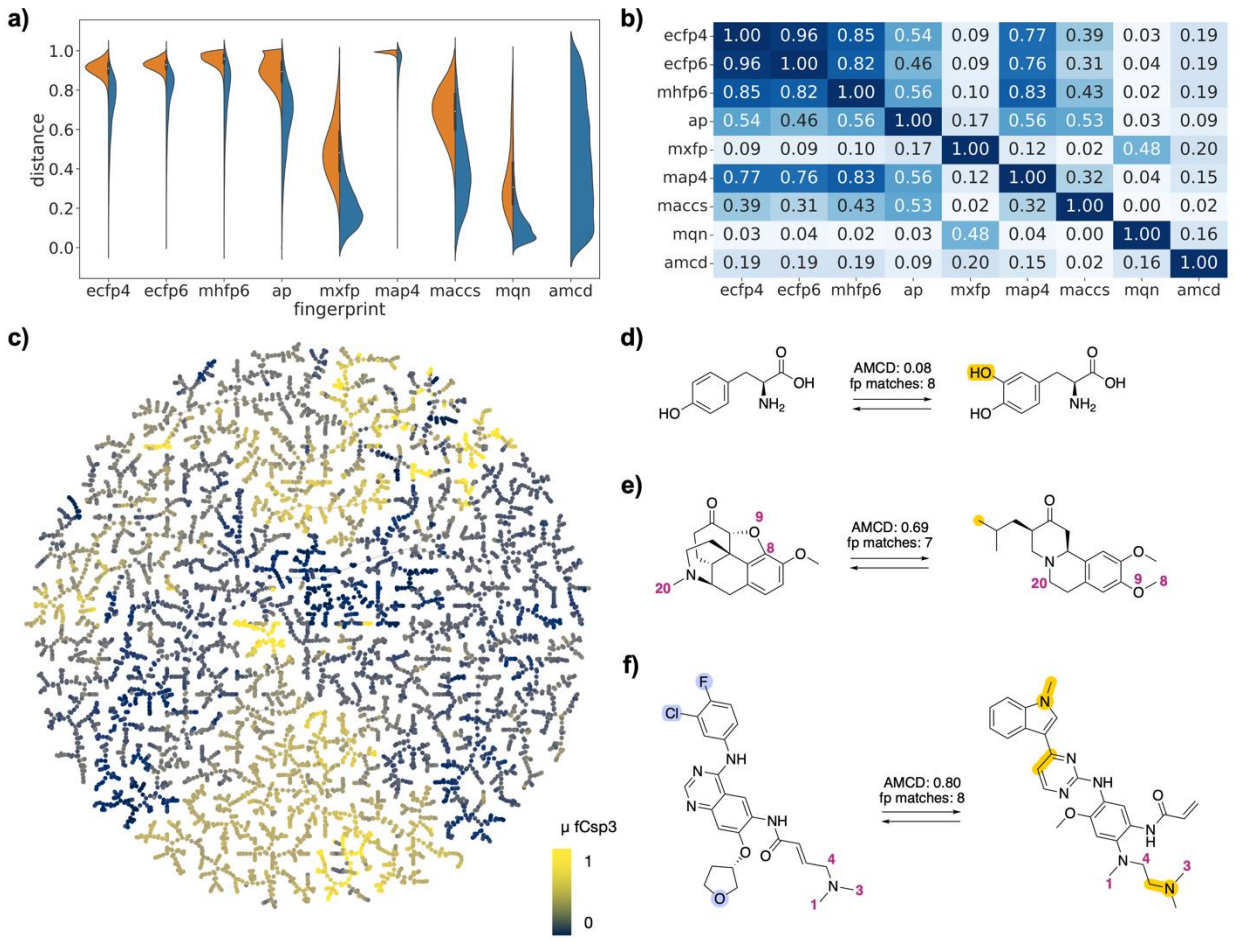


Figure 5. FDA-approved drugs as drug pairs. (a) Violin plot of dJ values in each of the fingerprints for all pairs (left, orange) or for selected pairs (right, blue), and for atom mapping confidence distance (amcd) of selected pairs (blue, last entry). (b) Heat map of correlation coefficients r2 between dJ values of different fingerprints, and between dJ values and amcd, calculated across all selected pairs. (c) TMAP of DRFP similarities for selected drug pairs. Each point is a different drug pair, color-coded by the fraction of sp3 atoms (Fsp3). See supporting information and https://tm.gdb.tools/map4/DRFP_FDA/ for additional color codes and for the interactive version of the map. (d) Atom-mapped drug pair L-tyrosine and L-DOPA related by a hydroxylation reaction. (e) Atom-mapped drug pair tetrabenazine and hydrocodone related by an alchemical double cyclization. (f) Atom-mapped drug pair afatinib and osimertinib related by a series of substituent and ring system changes. Atoms highlighted in blue are lost during the forward reaction, while atoms highlighted in yellow are gained. Interesting atom rearrangements as predicted by the RXNMapper are highlighted with their respective atom-mapping number. The full atom-mapping can be found in Figure A3.

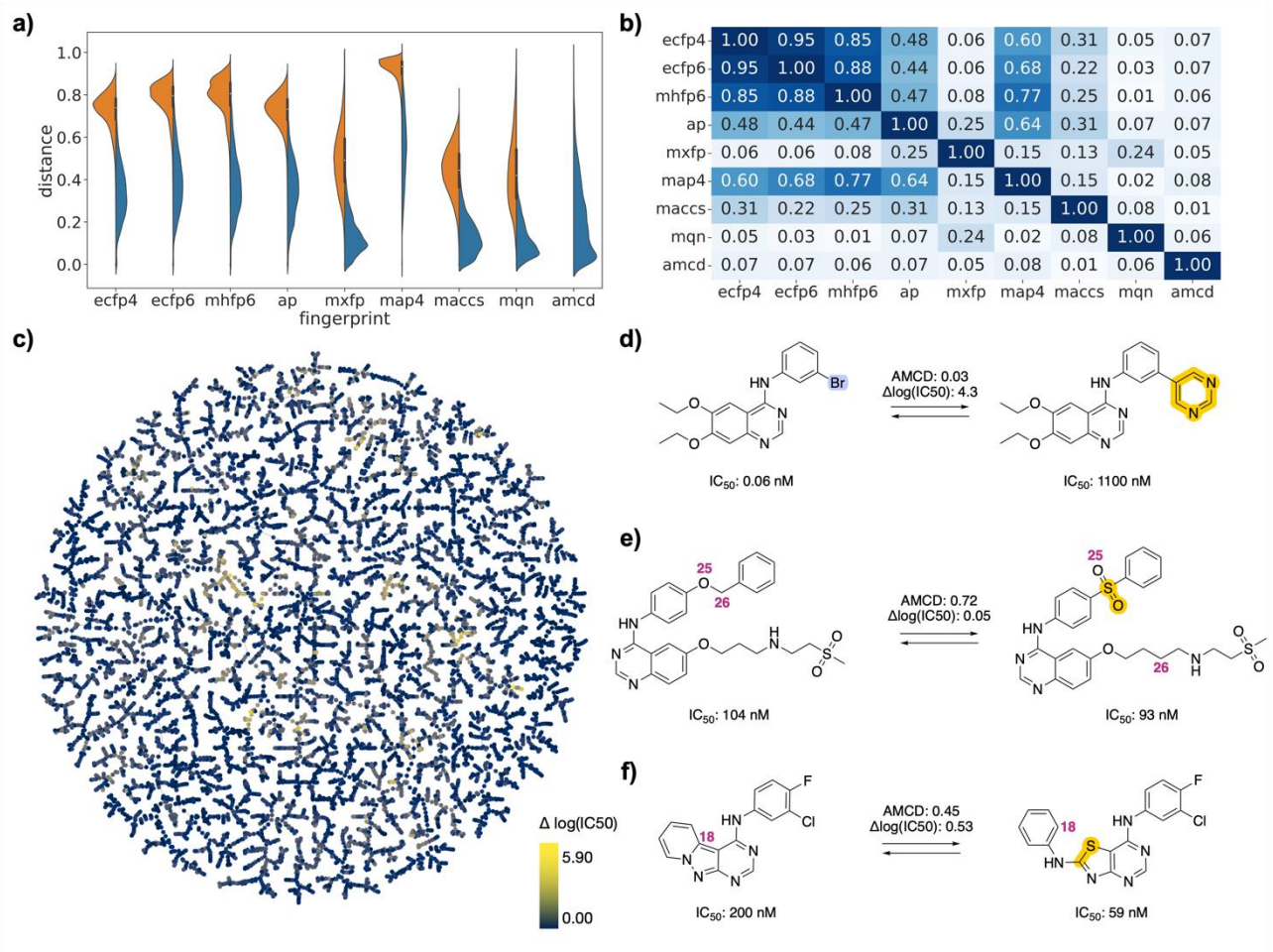


Figure 6. EGFR inhibitor drug pairs. (a) Violin plot of dJ values in each of the fingerprints for all pairs (left, orange) or for selected pairs (right, blue), and for atom mapping confidence distance (amcd) of selected pairs (blue, last entry). (b) Heat map of correlation coefficients r2 between dJ values of different fingerprints, and between dJ values and amcd, calculated across all selected pairs. (c) TMAP of activity differences. Each point is a different drug pair, color-coded by the activity difference. See supporting information and https://tm.gdb.tools/map4/DRFP_EGFR/ for additional color codes and for the interactive version of the map. (d) Atom-mapped drug pair CHEMBL35820 and CHEMBL126974 related by a Suzuki coupling resulting in an activity cliff. (e) Atom-mapped drug pair CHEMBL460732 and CHEMBL14952 related by an alchemical double linker exchange preserving activity (f) Atom-mapped drug pair CHEMBL469997 and CHEMBL181275 related by an alchemical scaffold hopping preserving activity. Atoms highlighted in blue are lost during the forward reaction, while atoms highlighted in yellow are gained. Interesting atom rearrangements as predicted by the RXNMapper are highlighted with their respective atom-mapping number. The full atom-mapping can be found in Figure A4.

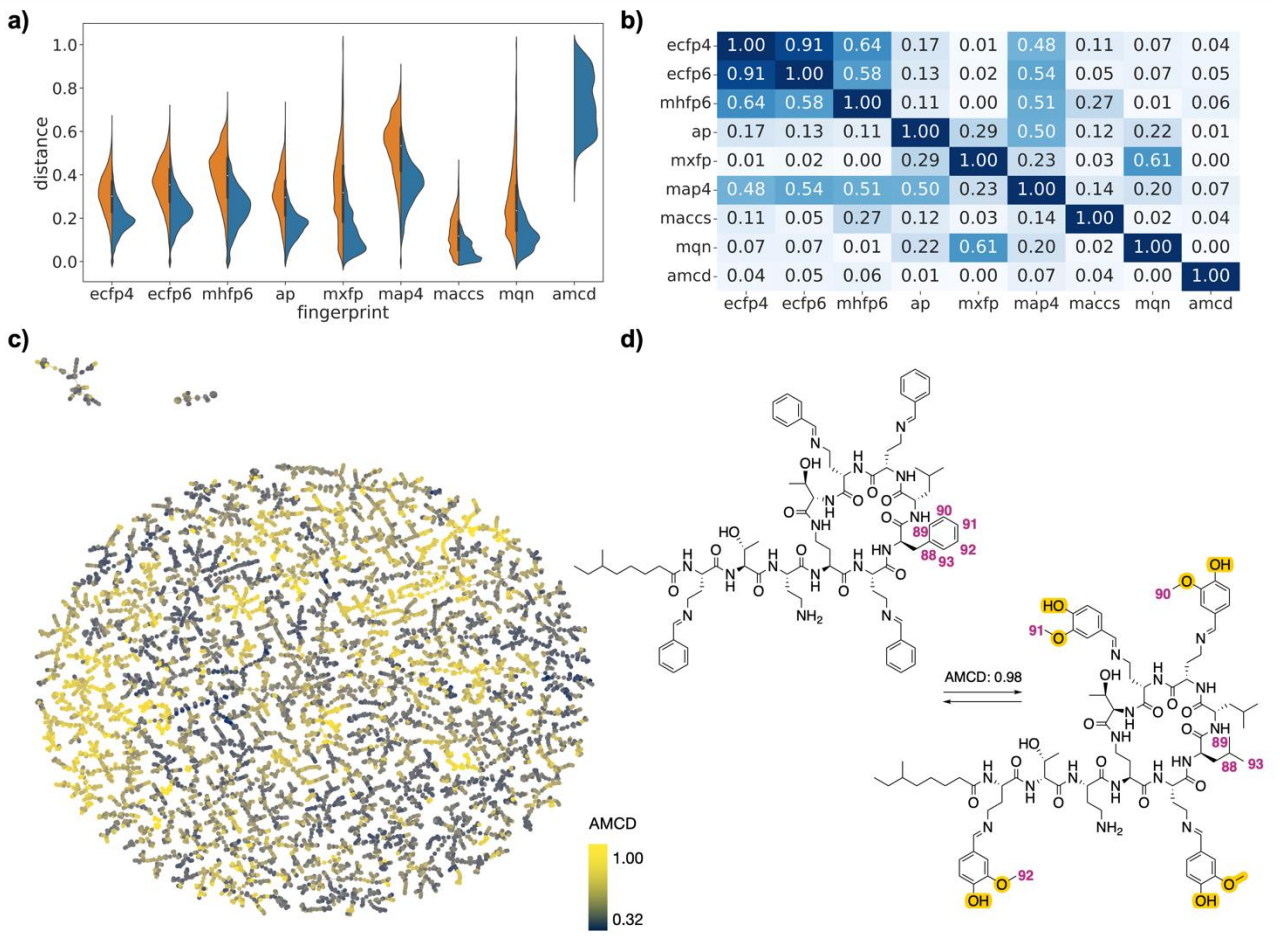


Figure 7. PMB analogs drug pairs. (a) Violin plot of dJ values in each of the fingerprints for all pairs (left, orange) or for selected pairs (right, blue), and for atom mapping confidence distance (amcd) of selected pairs (blue, last entry). (b) Heat map of correlation coefficients r2 between dJ values of different fingerprints, and between dJ values and amcd, calculated across all selected pairs. (c) TMAP of amcd values. Each point is a different drug pair, color-coded by the amcd value. See supporting information and https://tm.gdb.tools/map4/DRFP_PMB/ for additional color codes and for the interactive version of the map. (d) Atom-mapped drug pair CHEMBL1090265 and CHEMBL2372545 related by an imine exchange and a leucine→phenylalanine mutation. Atoms highlighted in blue are lost during the forward reaction, while atoms highlighted in yellow are gained. Interesting atom rearrangements as predicted by the RXNMapper are highlighted with their respective atom-mapping number. The full atom-mapping can be found in Figure A5.

4 One chiral fingerprint to find them all

This chapter is based on a scientific article previously published in the *Journal of Cheminformatics*. The article is reproduced here under the terms of the Creative Commons Attribution License (CC BY 4.0):

Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. *J Cheminform* **2024**, *16* (1), 53. https://doi.org/10.1186/s13321-024-00849-6.

Abstract

Molecular fingerprints are indispensable tools in cheminformatics. However, stereochemistry is generally not considered, which is problematic for large molecules which are almost all chiral.

Herein we report MAP4C, a chiral version of our previously reported fingerprint MAP4, which lists MinHashes computed from character strings containing the SMILES of all pairs of circular substructures up to a diameter of four bonds and the shortest topological distance between their central atoms. MAP4C includes the Cahn-Ingold-Prelog (CIP) annotation (*R*, *S*, *r* or *s*) whenever the chiral atom is the center of a circular substructure, a question mark for undefined stereocenters, and double bond cis-trans information if specified. MAP4C performs slightly better than the achiral MAP4, ECFP and AP fingerprints in non-stereoselective virtual screening benchmarks. Furthermore, MAP4C distinguishes between stereoisomers in chiral molecules from small molecule drugs to large natural products and peptides comprising thousands of diastereomers, with a degree of distinction smaller than between structural isomers and proportional to the number of chirality changes. Due to its excellent performance across diverse molecular classes and its ability to handle stereochemistry, MAP4C is recommended as a generally applicable chiral molecular fingerprint.

4.1 Introduction

Many computational tasks related to small molecule drug discovery, such as similarity searches,[23,125] target prediction, [149,151–154] ligand-based virtual screening[155] and visualization of large databases of drug-like molecules,[41,110,117,120,145,156–160] can be performed using vectors encoding molecular structure, called molecular fingerprints.[22,24] Remarkably, molecular fingerprints work quite well to classify and compare bioactive molecules without considering stereochemical information, which is somewhat surprising considering that biological matter is essentially chiral and stereo-defined at the molecular level,[161–163] but also reflects the fact one only rarely needs to distinguish between different stereoisomers of small molecule drugs, in part simply because many drug-like compounds are achiral.

In the context of developing computational tools for new modalities including beyond-Ro5 molecules,[164,165] in our case for peptides with variable chain topology and stereochemistry,[166-168] we have adapted molecular fingerprints based on atom-pairs[35–37,138] for large molecules such as peptides and proteins.[38,39,169] In particular, we combined atom-pair analysis and circular substructures as encoded the Morgan fingerprint ECFP4,[29,30] with the principle of data compression using MinHashing,[32,33,148,170] to design MAP4, a MinHashed Atom-Pair fingerprint. MAP4 encodes all possible pairs of circular substructures up to a diameter of four bonds in a molecule.[40] These pairs are written in the form of two canonicalized SMILES[171,172] separated by the shortest topological distance, counted in bonds, between the corresponding pair of central atoms. Remarkably, MAP4 distinguishes molecular structures across different compound classes spanning from small molecules to natural products, peptides and the metabolome, for which other fingerprints such as the classical Morgan (ECFP4)[29] and Atom Pair (AP)[35] fingerprints fall short. In addition, MAP4 outperforms these and many other fingerprints in virtual screening benchmarks for both small molecule drugs[24] and peptides.[40]

Similarly to commonly used molecular fingerprints however, MAP4 does not include stereochemistry (cis-trans double bonds, enantiomers and diastereomers), which is clearly an omission considering that most molecules beyond Ro5, such as diverse natural products and synthetic

compounds in the public databases ChEMBL,[142] COCONUT,[173] and ZINC,[141] are chiral (**Figure 8a**). To correct this omission and enable the cheminformatic analysis of compounds with multiple chiral centers such as carbohydrates and peptides, we now report MAP4C, an improved version of the MAP4 fingerprint. MAP4C includes the description of chiral centers following the Cahn-Ingold-Prelog (CIP) nomenclature in a fraction of molecular shingles (**Figure 8b/c**), as well as double bond stereochemistry.

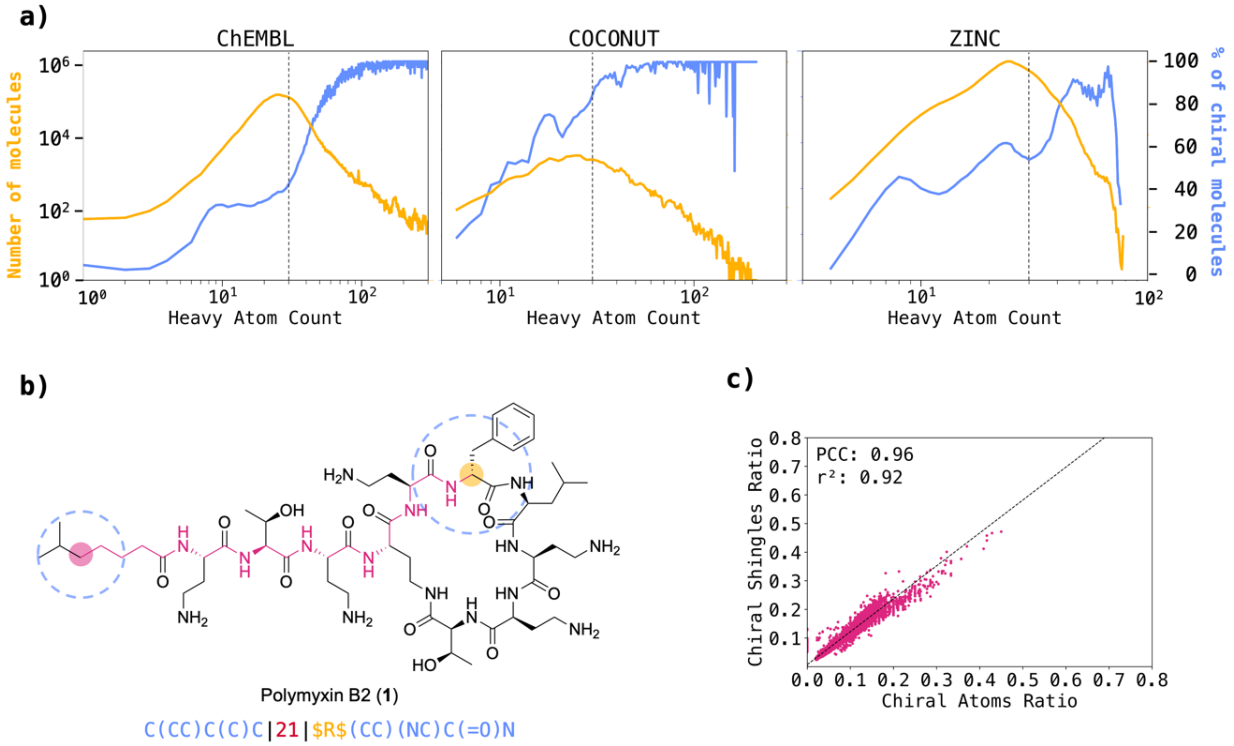


Figure 8. Molecular chirality and fingerprints. (**a**) Correlation between chirality and heavy atom count (HAC) across ChEMBL, COCONUT, and ZINC datasets. The blue line depicts the percentage of chiral molecules relative to HAC. A steady increase in the percentage of chiral molecules is observed with increasing HAC. The yellow line represents the total count of molecules corresponding to each HAC. (**b**) Chiral shingle generation concept exemplified on a selected atom pair of polymyxin B2. The generated shingle corresponds to the pair of circular substructures (blue) separated by the shortest topological distance (red) of their central atoms. Whenever the central atom of a substructure is chiral, the atom symbol in the substructure SMILES is replaced by the Cahn-Ingold-Prelog (CIP) descriptor (R, S, r, or s), or by a question mark (?) if the stereochemistry is not defined, bracketed by two "$" characters (yellow). (**c**) Percentage of molecular shingles containing chiral information vs. percentage of chiral atoms in the molecule for MAP4C (largest diameter of four bonds). These percentages were computed using a dataset of chiral molecules uniformly sampled from the Riniker & Landrum benchmark. The high r^2 and Pearson correlation coefficients underscore a strong association between the two variables.

4.2 Methods

4.2.1 Fingerprint design

The chiral version of the MinHashed Atom-Pair fingerprint (MAPC) was implemented in Python using RDKit following these steps:

1. At every non-hydrogen atom, extract all circular substructures up to the specified maximum radius as isomeric, canonical SMILES. Isomeric information ("@" and "@@" characters) is manually removed from the extracted SMILES, while the implicit E/Z-isomerism ("/", and "\" characters) are maintained. Allene chirality and conformational chirality such as in biaryls or in helicenes are not considered, as they cannot be specified in the SMILES notation. Radius 0 is skipped.

2. At the specified maximum radius, whenever the central atom of a circular substructure is chiral, replace the first atom symbol in the extracted SMILES with its Cahn-Ingold-Prelog (CIP) descriptor bracketed by two "$" characters ($CIP$). The CIP descriptor of the chiral atom is defined on the entire molecule, not on the extracted substructure.

3. At each radius, generate shingles for all possible pairs of extracted substructures. Each shingle contains two substructures and their topological distance in following format: "substructure 1 | topological distance | substructure 2".

4. MinHash the list of shingles to obtain a fixed sized vector. The MinHashing procedure is explained in detail in our previous publication.[32,40]

4.2.2 Benchmark

The virtual screening performance of the MAPC fingerprint was evaluated in a comparative study with commonly used fingerprints (ECFP4,[29] ECFP6,[29] Atom-Pair[35]) in a benchmark adapted from Riniker and Landrum.[24] Since the structure SMILES in the original benchmark do not contain any stereochemistry, the respective chiral SMILES (when applicable) were retrieved from the DUD,[174] MUV[175] and ChEMBL[142] databases using the provided compound IDs.

Additional 60 peptide sets were included in the benchmark to test the performances of the fingerprints for large biomolecules. For each of 30 random linear sequences, a set containing 10,000 single-point mutants and a set containing 10,000 scrambled versions of the random sequence were generated and BLAST analogues labelled as actives. The precise generation procedure of the peptide datasets is described in our previous publication.[40]

For every set, 5 randomly selected actives were extracted and stored in a separate file. The mean and standard deviation of pairwise ECFP4C Tanimoto and MAP4C Jaccard similarities of the five selected actives are reported in the **Figure B1-2**. Each of the selected actives was used as a query to rank the remaining compounds in the set based on fingerprint similarity (Jaccard similarity for MinHashed fingerprints; Dice similarity for folded fingerprints). AUC, EF1, EF5, BEDROC20, BEDROC100, RIE20 and RIE100 metrics were calculated for the obtained ranked lists and averaged along the 5 queries for every set in the benchmark. Additionally, the fingerprints were ranked based on the obtained performance metrics and finally the average rank of each fingerprint determined for all metrics. Pearson correlation coefficients and Friedman-Nemenyi post-hoc tests were calculated for all fingerprint pairs using the scipy and scikit-posthocs Python libraries.

4.2.3 Stereoisomers, isomers and scrambled sequences

We enumerated all possible stereoisomers of molecules **1 - 14** (**Figure 10**) by generating all possible isomeric SMILES combinations, canonicalizing them, and removing duplicates. We additionally enumerated all possible permutations of ln65 (**7**) and polymyxin B2 (**1**) sequences, obtaining a total of 330 and 1,512 scrambled sequences respectively. Structural isomers of 1,4-diaminocyclohexane

(**15)** and aminopiperazine (**16)** were extracted from GDB-13 using the MQN-browser.[176,177] The extracted sets contained 203 structural isomers of **15,** of which 156 contained one or more stereocenters and 48 structural isomers of **16,** of which 29 contained one or more stereocenters. For each structural isomer, all possible stereoisomers were generated using the RDKit "EnumerateStereoisomers" function, yielding 746 unique structures for **15** and 126 for **16**. For all stereoisomers and permutations, fingerprints were calculated as 2048-bit vectors.

4.2.4 TMAP

The indices obtained from the MAP4C calculation were used to create a locality-sensitive hashing (LSH) forest of 32 trees. For each molecular structure, the 500 approximate nearest neighbours in the MAP4C feature space were extracted from the LSH forest and used to calculate the TMAP layout.[41] The resulting layout was displayed in an interactive TMAP using the open-source Faerun package.[145]

4.3 Results and Discussion

4.3.1 Encoding stereochemistry in MAP fingerprints

The MAP (MinHashed Atom-Pair) fingerprint of a molecule consists in a series of MinHashes computed from the list of its molecular shingles.[32,33,148,170] A molecular shingle is written for each possible pair of circular substructures of a given diameter (2 bonds for MAP2, 4 bonds for MAP4, 6 bonds for MAP6), written as canonicalized SMILES, separated by the shortest topological distance separating the central atoms, counted in bonds.[40] We preserve the *Z/E* double bond information in all shingles whenever the entire double bond is included in a shingle. To encode stereocenter information into our fingerprints, we label chiral atoms with their Cahn-Ingold-Prelog (CIP) descriptor (*R*, *S*, *r* or *s*), as computed by RDKit, whenever stereochemistry is defined, or label them with a question mark ("?") if stereochemistry is not specified. Importantly, we only apply the chiral label when a chiral atom is the central atom of a circular substructure and only for shingles with the largest diameter considered. The concept is illustrated for one of the possible pairs involving the stereocenter in polymyxin B2 (**1**, (**Figure 8b**).

When applied to a dataset of chiral molecules uniformly sampled from the Riniker and Landrum benchmark (**Figure B1**),[24] we find that the percentage of molecular shingles containing chiral information is approximately the same as the percentage of chiral atoms in a molecule for MAP2C (largest diameter of two bonds, **Figure B2a**), MAP4C (largest diameter of four bonds, **Figure 8c**) and MAP6C (largest diameter of six bonds, **Figure B2b**). Most importantly, chiral information only appears in a relatively small fraction of all possible shingles, such that any defined stereoisomer of a molecule has a relatively high similarity to the molecule without assigned stereochemistry, for which the MAPC fingerprint is identical to the MAP fingerprint.

4.3.2 Virtual Screening Benchmark

The relevance of any molecular fingerprint for drug discovery can be tested by attempting to retrieve known bioactive compounds for a given target by nearest-neighbour searches from one of the known active compounds in a dataset in which the known actives have been mixed with so-called decoys. These decoys are molecules selected randomly from databases to have similar physicochemical properties as the actives, but which are not documented to be active on the target. Here we tested MAP4C with the reference benchmarking dataset of Riniker and Landrum for small molecule drugs,[24] which considers 118 active and decoy datasets taken from DUD,[174] MUV,[175] and ChEMBL.[142] For larger molecules, we used our previously reported set of 60 different randomly chosen 10-, 15- and 20-mer peptides mixed with either random single point mutants (30 sets), or sequence scrambled analog (30 sets),[40] for which we challenge the fingerprint to retrieve BLAST search analogs.[178]

Both of these benchmarks tested the ability of the fingerprints to retrieve bioactive analogs. Here we compared the performance of MAP2C, MAP4C, and MAP6C with their respective achiral counterparts, as well as with reference binary fingerprints ECFP4, ECFP6, and AP, and their corresponding chiral versions (ECFP4C, ECFP6C, and APC). The primary objective of the benchmark experiment was to ensure that the inclusion of chirality does not compromise the baseline virtual screening capabilities of the original MAP fingerprint. All fingerprints demonstrated comparable performances across various test sets and performance metrics, showing that including

chirality information was not detrimental to fingerprint performance in these benchmarks (**Figure 9a/b** and **Figure B5-9**).

Interestingly, the ranks of the different fingerprints for the various performances measures showed that the chiral MinHashed fingerprints were slightly ahead of the other fingerprints, with MAP4C appearing with the best ranks in the small molecule benchmark and MAP6C in the peptide benchmark (**Figure 9c**). We conducted a pairwise Friedman-Nemenyi test across all performance metrics to assess the statistical significance of performance differences among the various fingerprints (**Figure B10-16**). Performance differences within the same fingerprint groups (e.g., MAP and MAPC; ECFP and ECFPC; AP and APC) were not statistically significant, whereas performance differences between different groups typically were. Furthermore, the difference between chiral and non-chiral fingerprints were not significant, indicating that observed performance advantages were an artefact of rank combination rather than intrinsic differences.

However, MAPC fingerprints significantly performed better than ECFP(C) and AP(C) fingerprints (with exception of AP(C) for the AUC metric). MAPC fingerprints provide high local precision, akin to ECFPs, as well as global structure encoding, akin to AP fingerprints. This combination is particularly effective in scenarios where both local precision and global structure are relevant to differentiate between active and non-active molecules, possibly explaining the higher performance of the MAPC fingerprints compared to ECFPC and APC.

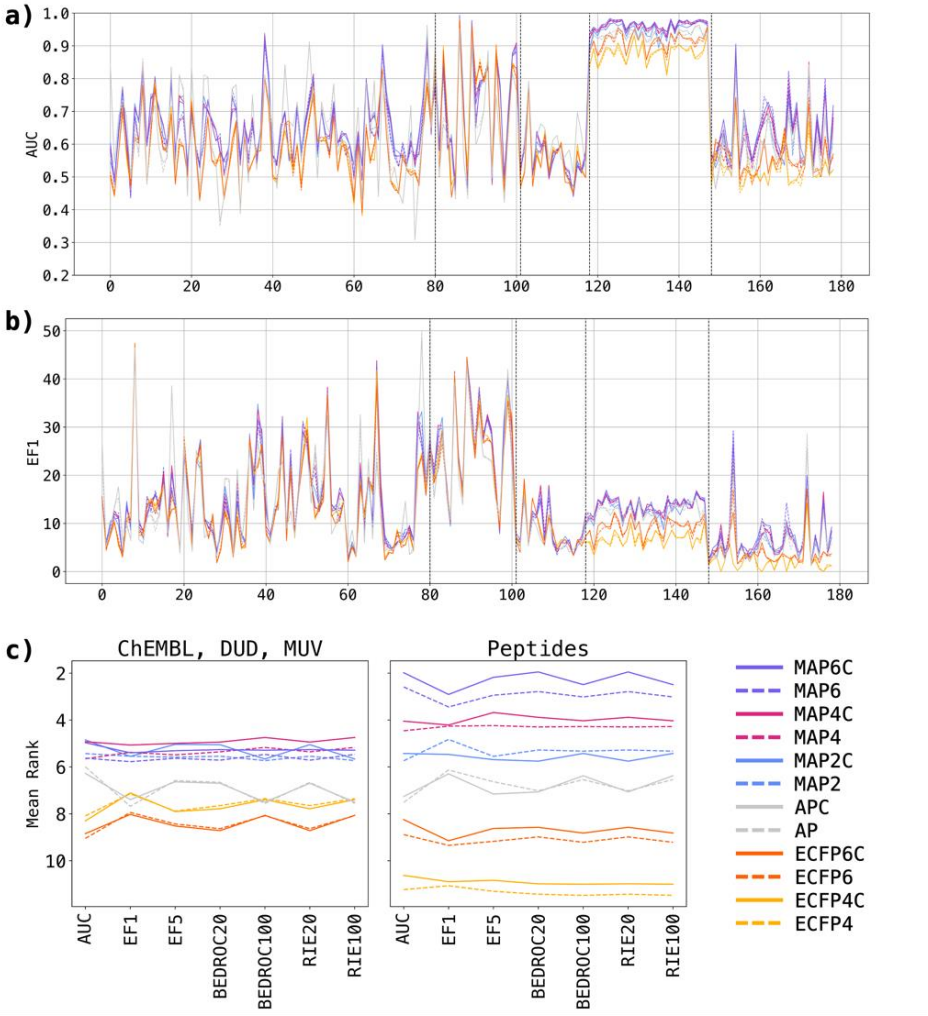


Figure 9. Virtual Screening benchmark a) AUC and b) EF1 of MAP6 (purple), MAP4 (magenta), MAP2 (blue), AP (grey), ECFP6 (orange) and ECFP4 (yellow) and across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs. c) Mean ranks of fingerprints across all virtual screening datasets for each metric. Small molecule sets (ChEMBL, DUD, MUV) and peptide sets are presented separately to highlight the differences in relative performance.

4.3.3 Finding all stereoisomers

In addition to be on par with non-chiral fingerprints for the above virtual screening benchmarks, one would expect a chiral fingerprint to distinguish all possible stereoisomers of a chiral molecule. To test the chiral differentiation of our fingerprints, we investigated their ability to assign a different fingerprint value for each stereoisomer on a series of stereochemically complex molecules comprising carbohydrates, peptides and macrocyclic natural products containing up to thousands of stereoisomers per molecule (**Figure 10** and

Table 1).

For carbohydrates, both MAP6C and MAP4C readily distinguished the 32 stereoisomers of α-D-glucopyranose (**2**), the 1024 stereoisomers of the disaccharide lactose (**3**), the 528 possible stereoisomers of the non-reducing, C_2-symmetrical α-diglucoside trehalose (**4**), the 16,384 stereoisomers of the aminoglycoside antibiotic validamycin A (**5**), and the nine possible stereoisomers of the signalling carbocyclic sugar *myo*-inositol (**6**). By contrast, the four other chiral fingerprints tested all fell short in at least one of the six cases, and APC failed on all of them.

Our MinHashed fingerprints performed very well with peptide stereoisomers. In the case of the antimicrobial undecapeptide ln65 (**7**), a membrane disruptive antimicrobial peptide whose activity/toxicity balance is modulated by stereochemical variations, and which motivated the present study,[168] the three chiral MAP fingerprints distinguished all the 2,048 possible stereoisomers. By contrast, ECFP6C only saw about half of them and ECFP4C and APC distinguished less than 10%, most likely because this peptide is composed of only lysine and leucine residues, which reduces the number of possible substructures. The chiral MAP fingerprints also distinguished the 330 possible sequence-scrambled isomers of **7** and the 675,840 possible stereoisomers of sequence-scrambled isomers of **7**. By comparison, APC succeeded for the 330 scrambled sequences but failed on the larger set, and both chiral ECFPs failed in both cases, which can be attributed to the absence of long-range substructures in ECFP fingerprints.

The ability of chiral MAP fingerprints to perceive peptide stereoisomers was also well illustrated by their ability to distinguish all 512 stereoisomers of the cell-penetrating peptide nona-arginine (**8**),[179,180] as well as the 4,096 stereoisomers of polymyxin B2 (**1**), used as last resort antibiotic against multidrug resistant bacteria.[97] In the latter case, our fingerprints also distinguished between the 1,512 possible sequence-scrambled isomers of **1**, the 774,144 possible sequence-scrambled stereoisomers of **1**, as well as between the 531,441 possible assignments of chirality as *R, S*, or undefined stereochemistry in the 12 chiral centers of **1**. An undefined stereochemistry corresponds to a stereorandomized position accessible by chemical synthesis using a racemic amino acid at that

position (stereorandomization at multiple position can lead to partially active analogs as reported for **1**).[181] In all of these cases, APC and ECFPCs were unable to distinguish all possibilities.

Macrocyclic natural products with rotational symmetries were particularly challenging for chiral fingerprints. For instance, only MAP4C and MAP6C correctly identified the 136 possible stereoisomers of the cyclic peptide antibiotic quinaldopeptin (**9**) and the 2,080 stereoisomers of the cytotoxic macrocyclic depsipeptide onchidin (**10**), two natural product macrocycles with C_2 symmetry. By contrast, the 528 stereoisomers of the C_2 symmetrical antimicrobial macrocyclic peptide gramicidin S (**11**) were only distinguished by MAP6C. Furthermore, none of the chiral fingerprints tested was able to cope with the C_3 symmetrical dodecadepsipeptide antibiotic valinomycin (**12**, 1,376 stereoisomers), the C_4 symmetrical macrolide ionophore antibiotic nonactin (**13**, 16,456 stereoisomers), or the C7 symmetrical hepta-arginine cyclic peptide NP213 developed as antifungal agent (**14**, 20 stereoisomers). Note that all fingerprints were used with 2,048-bits, but that performance did not increase significantly when using much larger bit sizes or without MinHashing or folding.

Table 1. Stereoisomer and scrambled sequence distinction task for selected natural products and peptides with multiple chiral centers and varying degrees of internal symmetry.

Query[a]	N / Sym.[b]	Total[c]	MAP6C	MAP4C	MAP2C	APC	ECFP6C	ECFP4C
α-D-glucopyranose (**2**)	5 /-	32	32	32	32	11	32	32
Lactose (**3**)	10 / -	1,024	1,024	1,024	992	443	1,024	1,024
Trehalose (**4**)	10 / C_2	528	528	528	516	336	528	512
Validamycin A (**5**)	14 / -	16,384	16,384	16,384	16,384	7,657	16,384	16,384
Inositol (**6**)	6 / C_{6v}	9	9	9	9	1	1	1
ln65 (**7**)	11 / -	2,048	2,048	2,048	2,048	196	1,140	36
ln65 (scrambled)	11 / -	330	330	330	330	330	8	4
ln65 (dia × scrambled)	11 / -	675,840	675,840	675,840	675,840	90,217	38,500	144
R_9 (**8**)	9 / -	512	512	512	512	146	88	12
Polymyxin B2 (**1**)[d]	12 / -	4,096	4,096	4,096	4,096	2,500	4,096	1,536
PMB2 (scrambled)[e]	9 / -	1,512	1,512	1,512	1,512	1,512	861	75
PMB2 (dia × scrambled)[f]	9 / -	774,144	774,144	774,144	774,144	287,631	602,003	9,312
PMB2 (*R, S* or undefined)	12 / -	531,441	531,441	531,441	531,441	277,901	531,441	137,781
Quinaldopeptin (**9**)	8 / C_2	136	136[g]	136	134	64	132	90
Onchidin (**10**)	12 / C_2	2,080	2,080	2,080	2,064	469	1,760	810
Gramicidin S (**11**)	10 / C_2	528	528	504	334	25	448	243
Valinomycin (**12**)	12 / C_3	1,376	1,250	714	416	112	616	27
Nonactin (**13**)	16 / C_4	16,456	16,425	16,176	10,045	13,189	6,474	675
NP213 (**14**)	7 / C_7	20	7	13	17	13	5	3

[a] Name and nr. of molecule. See Figure 4 for structural formulae. [b] N = number of stereocenters in the molecule. Sym. = rotational molecular symmetry for the molecule without chiral labels. [c] Number of possible stereoisomers considering inversion of all chiral centers in the molecule and the internal symmetry, or number of sequence isomers (scrambled). The number of different fingerprint values for each fingerprint type is given in the following columns. All fingerprint were used with 2,048 bit size unless otherwise noted. [d] all stereocenters in the molecule are considered. [e] amino acids are scrambled, the *N*-terminal fatty acid and the branching Dab residue are maintained. [f] only the α-carbon chirality of the scrambled residues was considered here, which corresponds to 512 stereoisomers per scrambled sequence. [g] with 4,096 bits, only 135 different FP values are obtained with 2,048 bits due to a bit collision.

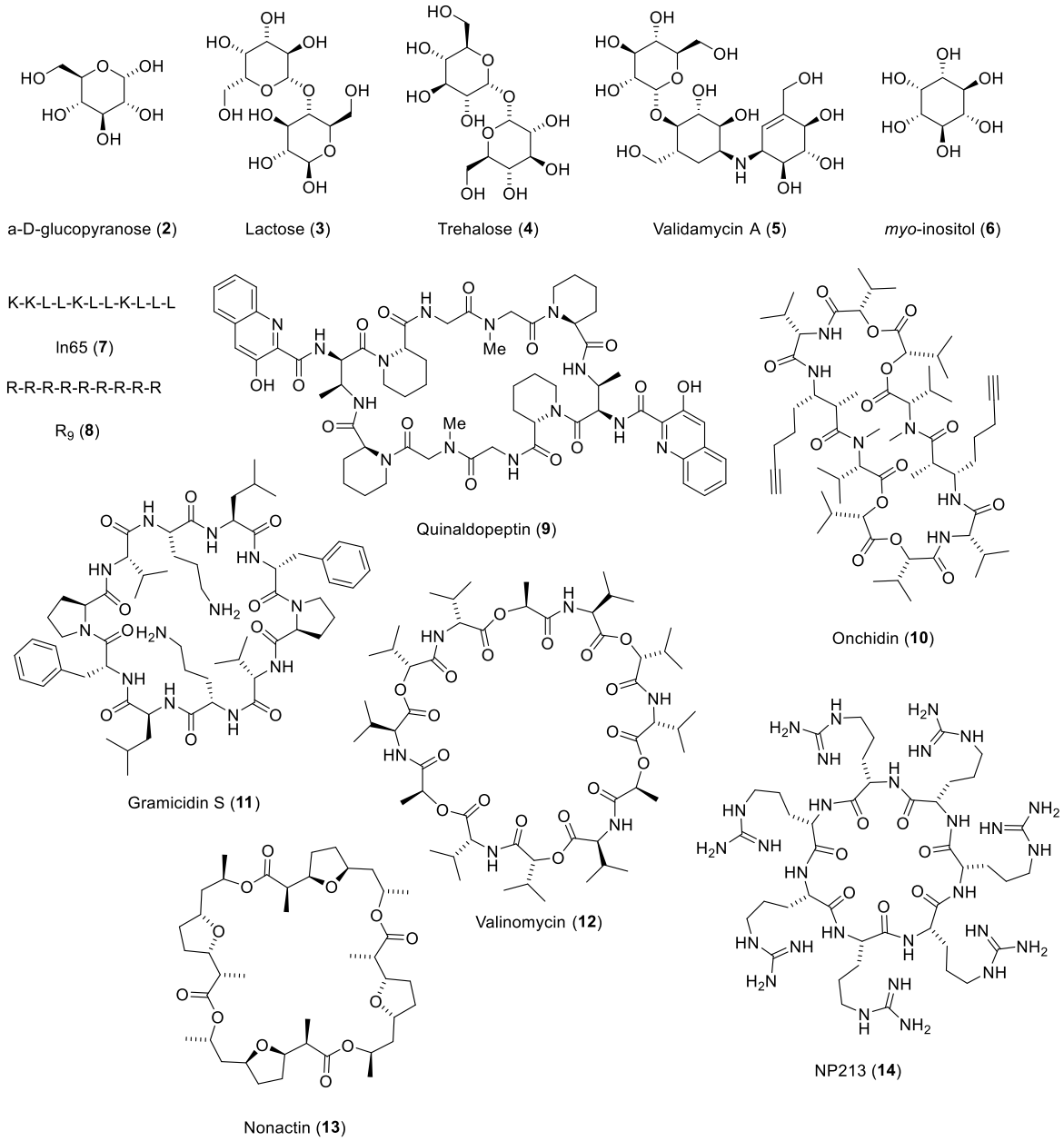


Figure 10. Structures of natural products and peptides selected for the stereoisomer distinction task.

4.3.4 Ranking stereoisomers versus isomers

The degree of differentiation between stereoisomers should be proportional to the number of stereochemical changes between any two stereoisomers and should also be smaller than the difference to a different molecule such as a structural isomer. We tested the ability of our chiral fingerprints for this task for small and large molecules separately. As a test case for small molecules, we computed Jaccard distances between all pairs involving the 203 structural isomers of 1,4-diaminocyclohexane (**15**), a ring fragment which is enriched in bioactive molecules from ChEMBL,[182,183] and between all

pairs of stereoisomers in the set. We similarly analysed all pairs involving the 48 structural isomers of 4-aminopiperazine (**16**), a similar drug scaffold, and the stereoisomeric pairs within the set. Generally, MAPC distances were higher than those of other fingerprints. This outcome is unsurprising, given that MAPC encodes a notably greater number of features, which also contributes to its high precision. In both test cases, all six fingerprints ranked pairs stereoisomers closer to each other than pairs of structural isomers (**Figure 11a/b**).

For peptides, we measured Jaccard distances between pairs of scrambled-sequence isomers versus pairs of stereoisomers with the same sequence for ln65 (**7**) and polymyxin B2 (**1**). For peptides, the degree of sequence similarity can also be measured by the Levenshtein distance, which represents the minimum number of mutations necessary to transform one sequence into another one, considering residue type changes, stereochemical inversions, insertions and deletions (**Figure 11c/d** and **(Figure B17/Figure** B18). Jaccard distances generally increased with increasing Levensthein distances for all fingerprints. Similar to small molecules, distances between peptide stereoisomers were smaller than between sequence isomers only for chiral MAP fingerprints and APC. However, chiral ECFPs assigned larger distances to stereoisomers than to sequence isomers, which probably relates to their inability to distinguish many pairs of sequence isomers. For both ln65 (**7**) and polymyxin B2 (**1**), the lower Jaccard distances between stereoisomers compared to sequence isomers was well visible in TMAP representations of each dataset constructed using MAP4C as similarity measure (**Figure 12a/b**).[41] In both cases, there was a complete separation between the 2,048/512 stereoisomers of the parent peptide and the 330/1,512 sequence isomers.

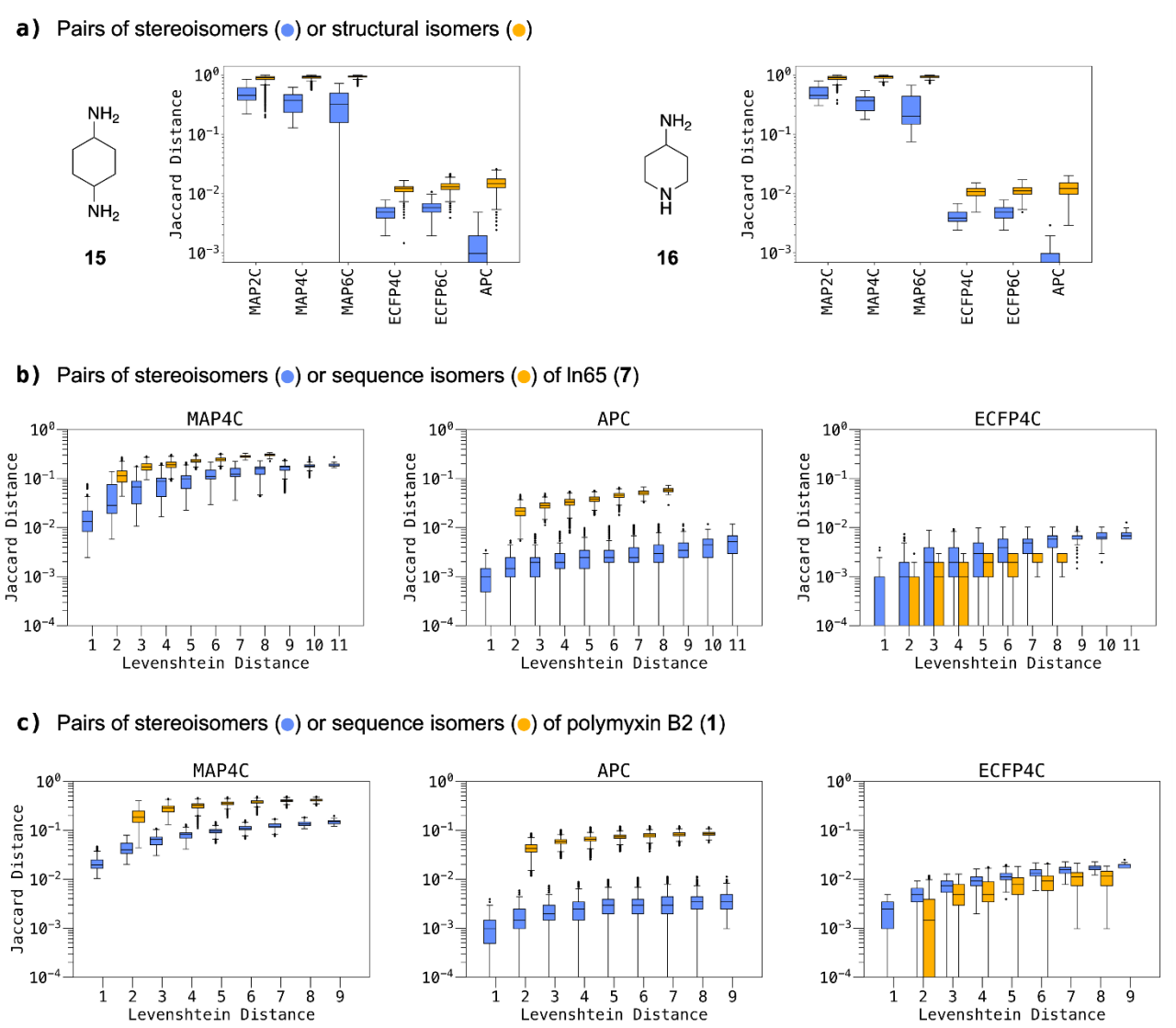


Figure 11. Differentiation between stereoisomers and structural isomers, shown as box plots of average Jaccard distances between pairs of stereoisomers (blue) or structural/sequence isomers (yellow). a) structural isomers of 1,4-diaminocyclohexane (203) and 4-aminopiperidine (48) and their diastereomers. The skewed distribution of Jaccard distance of 15 with MAP6C is caused by two outliers exhibiting a distance of 0 which cannot be represented on the log scale and is likely due to a bit-clash issue. b) sequence isomers (330) or diastereomers (2,048) of ln65 (7) as function of the Levenshtein distance separating each pair. c) sequence isomers (1,512) or diastereomers (512) of polymyxin B2 (1) as function of the Levensthein distance separating each pair. See Figure B10 and Figure B11 for plots with MAP6C, MAP2C and ECFP6C. See methods for details.

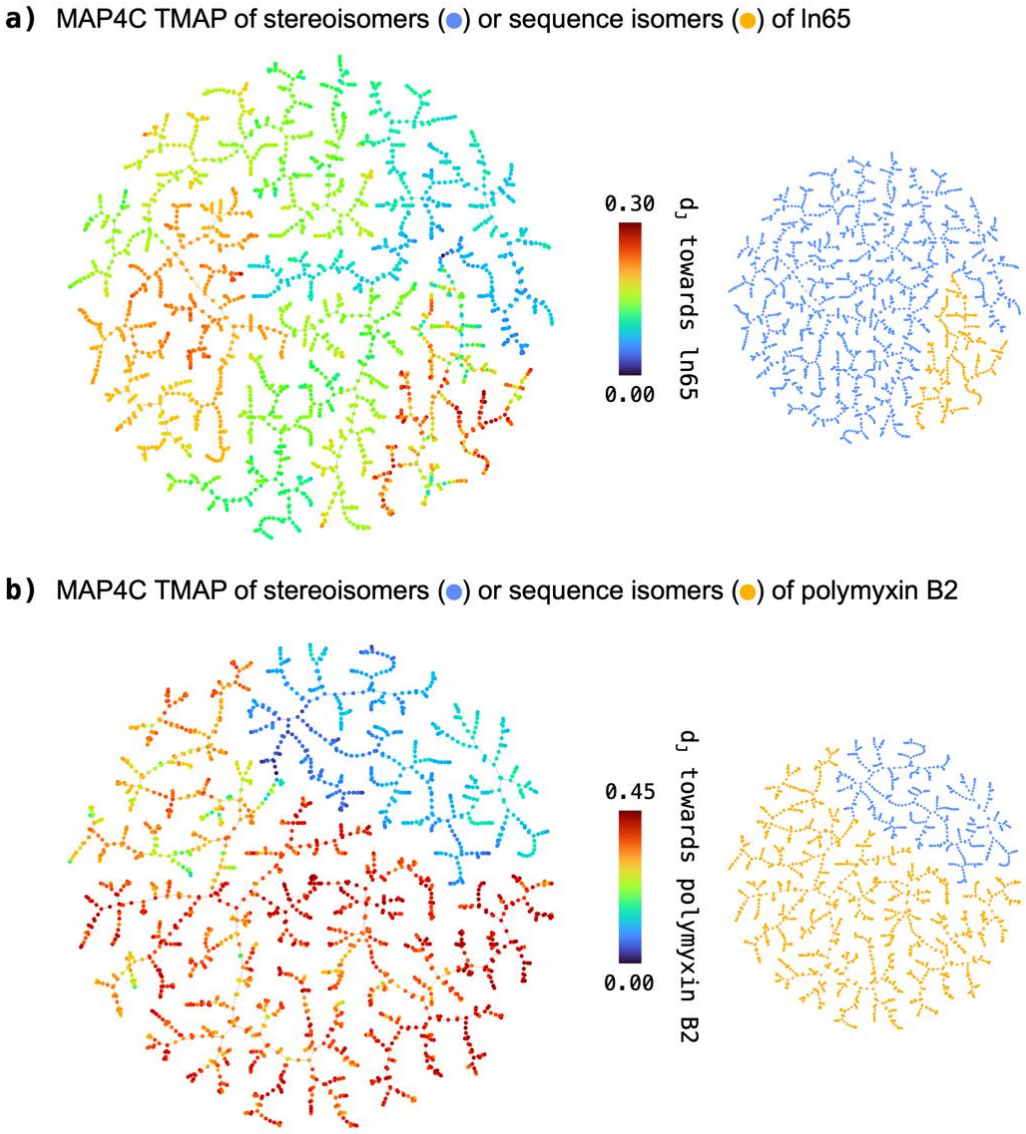


Figure 12. MAP4C TMAPs showing the Jaccard distance (d_J; rainbow) of stereoisomers (blue) and sequence isomers (yellow) towards their respective queries: (a) ln65, 2,048 diastereomers and 330 sequence isomers. The interactive version of the TMAP is accessible under https://tm.gdb.tools/map4/MAP4C_ln65/ (b) polymyxin B2, 512 diastereomers and 1,512 sequence isomers. The interactive version of the TMAP is accessible under https://tm.gdb.tools/map4/MAP4C_pmb2/.

4.4 Conclusion

In summary, the data above shows that the chiral versions of MAP fingerprints reported here perform as good as their achiral versions in non-stereoselective virtual screening benchmarks. Remarkably, our chiral MAP fingerprints are able to distinguish stereoisomers even in cases involving up to thousands of stereoisomers where the chiral versions of ECFP and AP do not perform well. Furthermore, the chiral MAP Jaccard distances between enantiomers or stereoisomers are generally shorter than for structural isomers, allowing to use chiral MAP fingerprints as a refinement of their achiral version. Because MAP4C computes faster than MAP6C due to the small number of atom pairs considered, we recommend MAP4C as the molecular fingerprint of choice for comparing molecules spanning from small drug-like building blocks to large natural products and peptides. The ability of our chiral fingerprint MAP4C to handle stereoisomers from small molecules to large natural products and peptides is unprecedented and opens the way for cheminformatics to include stereochemistry as an important molecular parameter across all fields of molecular design.

4.5 Declarations

4.5.1 Funding

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076).

4.5.2 Availability of data and materials

The source codes and datasets used for this study are available at https://zenodo.org/records/10389905 The code for MAPC can be found at https://github.com/reymond-group/mapchiral.

4.5.3 Competing interests

The authors declare that they have no competing interests.

4.5.4 Author contributions

MO designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper. Both authors read and approved the final manuscript.

5 Navigating a 1E+ 60 chemical space of peptide/peptoid oligomers

Orsi, M.; Reymond, J. Navigating a 1E+60 Chemical Space of Peptide/Peptoid Oligomers. *Molecular Informatics* **2024**, e202400186. https://doi.org/10.1002/minf.202400186.

Abstract

Herein we report a virtual library of 1E+60 members, a common estimate for the total size of the drug-like chemical space. The library is obtained from 100 commercially available peptide and peptoid building blocks assembled into linear or cyclic oligomers of up to 30 units, forming molecules within the size range of peptide drugs and potentially accessible by solid-phase synthesis. We demonstrate ligand-based virtual screening (LBVS) using the peptide design genetic algorithm (PDGA), which evolves a population of 50 members to resemble a given target molecule using molecular fingerprint similarity as fitness function. Target molecules are reached in less than 10,000 generations. Like in many journeys, the value of the chemical space journey using PDGA lies not in reaching the target but in the journey itself, here by encountering non-obvious analogs. We also show that PDGA can be used to generate median molecules and analogs of non-peptide target molecules.

5.1 Introduction

Since the advent of combinatorial chemistry in the early 1990's, which was triggered by the invention of the split-and-mix method yielding one-bead-one-compound libraries of millions of peptide and peptide-like oligomers in a few tens of synthetic operations,[193-195] drug discovery has been fascinated and partly driven by large numbers.[12,13,187] Approaches ranged from the "needle in a haystack" method of high-throughput screening typical for genetically encoded display libraries[188,189] and DNA-encoded libraries,[190,191] to the concept of chemical space guiding the design of focused libraries of small drug-like molecules,[109,192,193] fragments[183,194] and peptides.[166,195,196] Many projects are currently exploiting "make-on-demand" virtual libraries of a few billion members obtained by using various coupling chemistries to combine two to four building blocks, each being taken from a pool of thousands of building blocks, to form linear, branched or cyclic oligomers.[208-211] Despite of being rather constrained, this oligomer chemical space has proven amenable to virtual screening and sufficiently diverse to solve most drug discovery problems,[212-214] probably because biomolecules are themselves oligomers and their binding sites are usually suitable for partly flexible, pearl-string like molecules.[215-217]

Following up on our interest for exhaustive enumeration of chemical space, [182,207,208] here we aimed to extend the oligomer chemical space to reach up to a virtual library size of 1E+60, a common estimate for the total size of the drug-like chemical space.[13,14] We also aimed to demonstrate virtual screening at that library size focusing on ligand-based virtual screening (LBVS).[209,210] LBVS consists in identifying analogs of a reference bioactive compound by scoring the virtual library using molecular similarity measures such as molecular fingerprints,[22,24,155,211] or shape-based comparisons.[36,124,138,212,213]

As discussed below, we achieved our goals for the case of mixed peptide-peptoids potentially accessible by solid-phase peptide synthesis (SPPS),[214] moving up to 30-mers with 100 different building blocks to reach the required library size. To demonstrate LBVS, we modified our recently reported peptide design genetic algorithm (PDGA),[102] which evolves analogs of any target molecule

by performing mutations/selection cycles on sequences encoding a topologically diverse oligomer space using molecular fingerprint similarity as fitness function, an approach which is related to small molecule design genetic algorithms.[215,216] PDGA can be used to design new analogs of known peptides as recently demonstrated experimentally for antimicrobial peptide dendrimers.[217] Specifically, we computed the fitness function using the macromolecule extended atom pair fingerprint (MXFP)[38,39] and the chiral MinHashed atom pair fingerprint (MAP4C),[40,42] both designed for large molecules.

5.2 Methods

5.2.1 Building Blocks

Our set of 100 building blocks includes the 20 proteinogenic amino acids, their D-enantiomers, 12 further amino acids, 46 peptoids (*N*-substituted glycines)[218] as well as GABA and β-alanine, all available commercially or easily accessible in protected form for Fmoc-SPPS or for the submonomer synthesis method for peptoids (**Figure C1**).[219,220] To further augment diversity, we allowed 11 different acyl group to cap the *N*-termini, and allowed a single cyclization either via a cystine bridge or by amide bond formation between the C-terminus and the *N*-terminus or a primary amine side chain (at lysine and related diamino acids). All building blocks are encoded in SMILES notation, ensuring that their concatenation always leads to a valid molecule. Additionally, sequences are represented in linear format to facilitate mutation and cross-over operations within the genetic algorithm. In this format, "BBXXX" denotes a building block containing an amine and carboxylic acid, "bXXX" a diamino acid for sequence branching, "c" a C-to-N cyclization, “s” a cysteine for disulfide bridges, and "TXXX" an *N*-terminal cap**. Figure C2** illustrates the enhanced sequence format. Both, the enhanced sequence format, and the corresponding SMILES, are stored in the results files.

5.2.2 Genetic Algorithm

We modified our previously reported PDGA[102] by computing fitness functions either as the Jaccard distance (d_J) to the target molecule computed using the molecular fingerprint MAP4C,[42] saving all generated molecules at each generation as trajectory molecules, or as the City Block Distance (d_{CBD}) to the target molecule computed using the most recent version of MXFP,[39] here saving only molecules with $d_{CBD} \leq 300$ as trajectory molecules, a threshold which only retains molecules with a significant degree of similarity to the target. Each PDGA run was started either from 50 random linear sequences generated using the 100 available building blocks, or from 50 repetitions of a selected starting sequence (for traversal runs) and stopped either when the target was found or after 10,000 generations. For all runs, a mutation rate of 0.5, population size of 50 and free topology exploration were employed during the genetic optimization process. In each iteration, the 15 sequences nearest to the query are chosen as parents and mutated to create 35 new sequences, which are then added to the population. Mutation types include point mutations, deletions, insertions and cross-over. A second set of topology-changing mutations were added to the pool of possible mutations in the PDGA. These include forming and breaking of *C*-to-*N*-cyclizations, forming and breaking of branching points using diamino acids as well as forming and breaking of disulfide bridges by insertion of two cysteines.

5.3 Results and Discussion

5.3.1 A 1E+60 combinatorial library from 100 building blocks up to 30-mers

Due to its size, a chemical space of 1E+60 cannot be explicitly enumerated, leaving a formal combinatorial enumeration as the only viable option. Assembling N building blocks to form an oligomer of length M results in N^M possibilities, hence 1E+60 is readily reached in a 60-mer peptide using only 10 different amino acids, in line with the well-known combinatorial explosion of possibilities in peptide and protein sequences. However, reducing length M in the direction of small molecules requires an exponentially increasing number of building blocks N, for instance including all 20 proteinogenic amino acids would still require a 46-mer to reach 1E+60, and reducing oligomer

length to a tetramer assembly typical of small molecules would require 1E+15 building blocks, well beyond the known small molecule chemical space (**Table 2,** 2nd column).

Here we settled for 100 building blocks, reaching 1E+60 with a 30-mer, which lies within the size range of peptide drugs such as the HIV membrane fusion inhibitor enfuvirtide (34 residues)[221] or the diabetes/obesity drug semaglutide (31 residues).[222] To reach N = 100, we considered the 20 proteinogenic amino acids in L- and D- enantiomeric forms, together with simple non-proteinogenic amino acids as well as peptoids (*N*-alkylated glycine),[218] which can be easily assembled by SPPS with the sub-monomer approach.[223] All 100 building blocks selected were commercially available or easily accessible in a protected form suitable for peptide and/or peptoid submonomer SPPS (**Figure C1).**

Table 2. Influence of oligomer length M and number of building blocks N on virtual library size.[a)]

oligomer length (M)	Number of building blocks (N) required to reach N^M = 1E+60	Library size at length M with N = 100
60	10	1E+120
46	20	1E+92
30	**100**	**1E+60**
29	117	1E+58
15	10,000	1E+30
8	31,622,777	1E+16
4	1E+15	**100,000,000**

a) For linear amide-bond connected oligomers. Since we consider single strands and the amide bond is directional (CO-N is not equivalent to N-CO), there are no symmetrical sequences. In the present report library size is further increased by cyclization and diverse N-terminal caps (see text and methods).

With these 100 building blocks at hand, a virtual combinatorial enumeration of 1E+60 sequences was possible. To increase diversity, we allowed for eleven different *N*-terminal carboxylic acids, in

particular fatty acids as found in peptide antibiotics such as polymyxin[97] and which favour cellular uptake in natural products[224] and extend peptide circulation times via albumin binding.[225] We also added several options for cyclization to increase diversity (see methods for details). While these additional variations enlarged library size, it should be noted that library size depended primarily on oligomer length. For instance, reducing length by one unit to 29-mers reduced library size by 100-fold, implying that 99% of the library resided with 30-mers. Nevertheless, with 100 building blocks the virtual library still contained 100 million members for tetramers, well in the size range of the public archive PubChem (**Table 2**, 3rd column).[226]

5.3.2 Ligand-based virtual screening by genetic algorithm guided navigation

Virtual screening consists in computationally evaluating a dataset to select a restricted number of molecules for closer inspection. Here we used LBVS aiming to select analogs of a target compound by using a genetic algorithm approach with PDGA (**Figure 13**).[102] Genetic algorithms evolve a population for fitness by rounds of mutations and selection. In the context of our 1E+60 chemical space, this approach corresponds to a targeted navigation guided by the fitness function, which circumvents the need for evaluating every library member. We set out to test whether our PDGA would find its way through our 1E+60 virtual library, drawing from the selected set of 100 peptide/peptoid building blocks rather than only 20 amino acids to generate mutants.

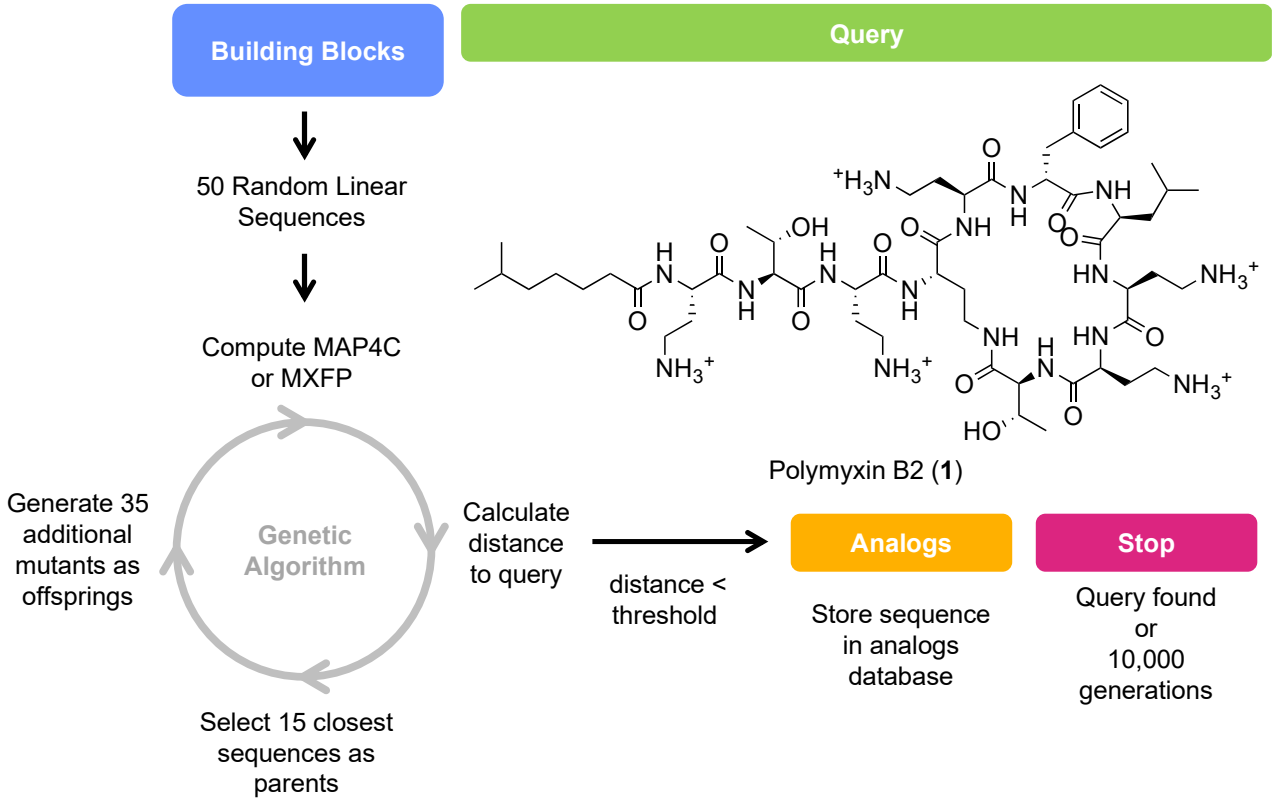


Figure 13. Design of PDGA. PDGA uses a list of input building blocks to generate a set of random linear sequences. The sequences are encoded using either the MAP4C or MXFP fingerprints. The fingerprints are used to determine the fitness of the sequences by calculating the distance towards a specified query molecule. Sequences with distances below a set threshold are stored in an analogs database. The 15 fittest sequences undergo rounds of mutations and crossovers in which building blocks and topology are changed to add 35 new sequences to the population. This process iterates until either the query is found or the PDGA reaches 10,000 generations.

We challenged PDGA to identify analogs of six known bioactive linear and cyclic peptides of various length in our 1E+60 library. The test cases were polymyxin B2 (**1**, 10 residues, antimicrobial),[97] gramicidin S (**2**, 10 residues, antimicrobial),[227,228] the mixed peptide/peptoid hybrid EB9 (**3**, 11 residues, antibacterial),[214] oncocin (**4**, 19 residues, antimicrobial),[229] cathelicidin BF (**5**, 30 residues, immunomodulatory peptide),[230] and circulin D (**6**, 30 residues, anti-HIV)[231] (**Figure 14**). In each case, we performed three PDGA runs of maximum 10,000 generations starting from 50 random sequences using the chiral fingerprint MAP4C, which encodes pairs of circular substructures with high precision including chirality.[40,42]

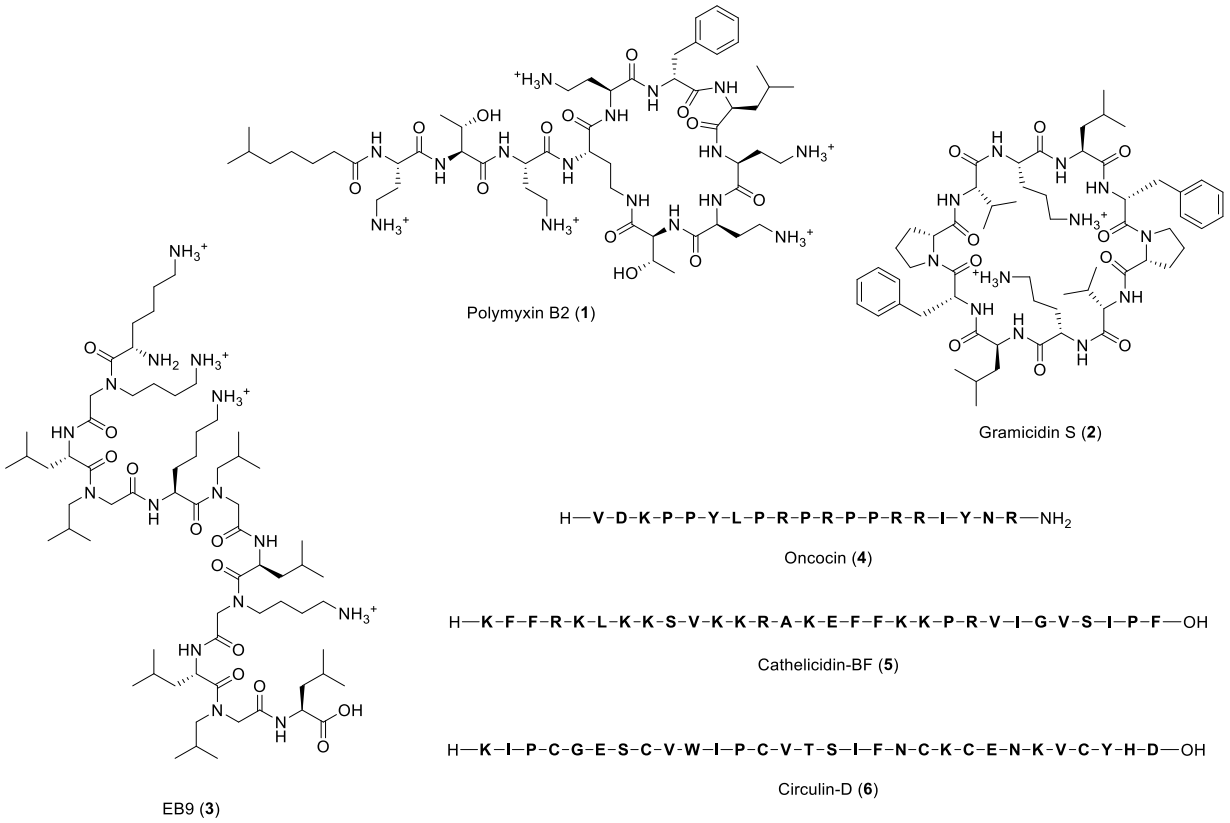


Figure 14. Structures of the selected queries for the PDGA runs using the MAP4C similarity as fitness function. The linear sequences (4, 5 and 6) are written with standard one-letter code for amino acids, with free N-terminus marked as "H-" and C-terminus in acid form "-OH" or amide form "-NH_2".

PDGA identified the target molecule in less than 10,000 generation in at least one of the three runs for each of these six peptides, including the two 30-mer peptides **5** and **6,** which required exploration of the full 1E+60 chemical space (**Table 3**). Since each generation only amounted to 35 new molecules, which were evaluated against the 15 best scoring molecules of the previous generation used as parents, the cumulative number of molecules generated in each trajectory only amounted to a few thousands, which is remarkably low considering the size of the explored chemical space. Note that the number of molecules per trajectory was approximately 30% lower when excluding stereoisomers. The presence of stereoisomers in the trajectory resulted from the presence of D- and L- residues in the building block set and the ability of MAP4C to rank each stereoisomer differently. Among the generated structures, PDGA delivered thousands of virtual screening hits characterized by a high similarity (Jaccard distance $d_J < 0.5$) to the target peptide.

The evolution of the best score (d_J to target) per generation as function of generation number illustrated how PDGA reached each target (**Figure 15** and **Figure C3**, upper row). After an initial round of approximately 10 generations, the best score started to decrease, indicating that the algorithm had found a way towards the target. After approximately 1,000 generations, the score had either decreased to zero and the target had been found, or the algorithm was stuck at an intermediate score. In terms of the cumulative number of new molecules generated, the increase per generation was approximately steady until the target had been found (**Figure 15** and **Figure C3**, lower row). When the target was not found however, the algorithm was unable to generate any new structures, indicating that the same 15 top scoring molecules kept being selected as parent in each round and that none of their mutants led to any improvement in the score, implying that a local minimum had been reached. Because the computational expense of correcting this limitation by introducing a duplicate molecule check at every iteration was found to be far too large and the target was usually found by repeating the run several times, the algorithm was not modified.

Table 3. Results of three parallel PDGA runs for queries 1-6.

Query	length	Structure[a]	# generations to query[b]			# unique structures (% with $d_J < 0.5$)[c]			# unique structures not counting diastereomers (% with $d_J < 0.5$)[c]		
			Run 1	Run 2	Run 3	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Polymyxin B2 (**1**)	10	cyclic peptide	894	1,371	>10k	6,934 (67)	6,362 (74)	7,877 (24)	5,123 (57)	4,792 (67)	4,851 (13)
Gramicidin S (**2**)	10	cyclic peptide	512	736	>10k	4,119 (69)	5,438 (80)	4,142 (13)	3,384 (63)	4,505 (76)	2,958 (9)
EB9 (**3**)	11	peptoid	2,485	2,295	>10 k	20,998 (36)	20,377 (44)	7,160 (32)	16,705 (32)	16,333 (41)	5,720 (28)
Oncocin (**4**)	19	linear peptide	5,350	5,629	>10k	46,591 (80)	39,835 (77)	55,462 (67)	22,023 (65)	27,829 (70)	32,698 (52)
Cathelicidin BF (**5**)	30	linear peptide	9,355	8,521	>10k	88,738 (86)	86,265 (87)	31,301 (86)	57,367 (81)	63,374 (83)	20,831 (80)
Circulin D (**6**)	30	Cyclotide[d]	8,133	>10 k	>10 k	73,535 (73)	37,526 (74)	33,738 (61)	43,550 (58)	23,368 (61)	26,092 (53)

a) see supporting information Figure S2 for structural formulae. b) number of generations used by PDGA to reach the query molecule. >10k indicates that the target was not found within 10k generations. c) d_J refers to the Jaccard distance calculated using MAP4C fingerprints. d) PDGA was run on the linear sequence lacking the cystine bridges.

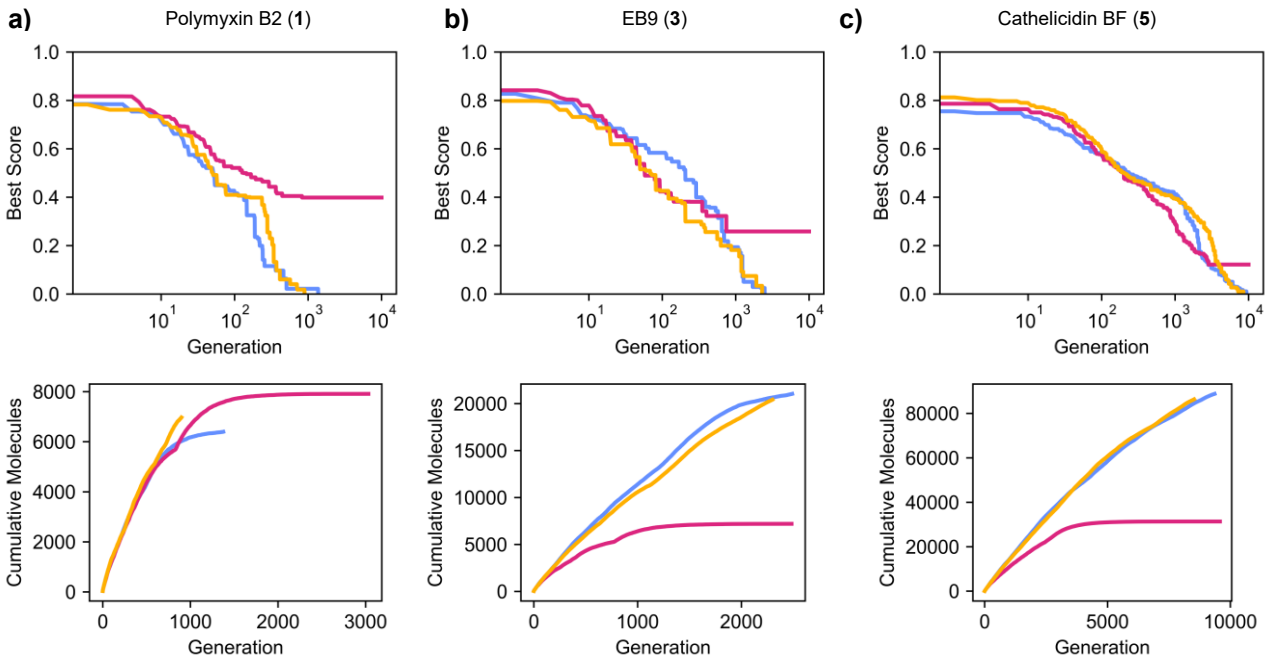


Figure 15. Analysis of three parallel PDGA runs starting from 50 random sequences towards selected queries. Top plots show the overall best score throughout the trajectory; the bottom plots show the cumulative number of unique new molecules generated throughout the trajectory for a) polymyxin B2, b) EB9, and c) cathelicidin BF. The best score refers to the MAP4C d_J of the closest structure generated up to that generation relative to the target.

To get a closer insight into the analogs (MAP4C $d_J < 0.5$) generated by PDGA, we focused on the case of polymyxin B2 (**Figure 16**). We compared the three PDGA runs with an additional self-run, starting PDGA from polymyxin B2 and letting the algorithm complete 10,000 generation independent of target identification. This self-run quickly exhausted itself and produced 1,906 unique analogs, significantly less than the approximately seven thousand analogs obtained for each PDGA run. Interestingly, each of the three runs produced a different set of analogs (**Figure 16a**). While it is not surprising that all 7,877 molecules in the failed run were unique to this run since it failed to converge on the target, the two successful runs only shared three common molecules and less than 100 with the self-run, although all molecules in these runs were highly similar to polymyxin B2, with an average Jaccard distance below 0.35 (**Figure 16b**). We also analysed the average number of mutations from the target using Levenshtein distance as a proxy. Analogs of the successful runs were on average three mutations away from the target, while the self-run only produced point mutants and molecules from the failed run remained approximately 9 mutations away from polymyxin B2 (**Figure 16c**).

A closer analysis of the successful runs revealed that many analogs combined multiple mutations with a high similarity to the target, as exemplified with analog **7** (**Figure 16d**). Such analogs are particularly interesting since they would be difficult to identify without PDGA compared to single point mutant from the self-run, which do not require an algorithm for design. When displayed on a tree-map (TMAP)[41] computed using MAP4C similarities, molecules from the two successful runs and the self-run were intermixed, indicating that they occupied a similar chemical space. Note however that two clusters of molecules from Run 1 (blue) or Run 2 (yellow) were visible, which contained early generation molecules with high Jaccard distance. Molecules from Run 3, which did not reach the target, also remained at high Jaccard distance and occupied a separate area of the map, reflecting their very different structural type, which featured a large, unbranched macrocycle exemplified by analog **8** (**Figure 16d**).

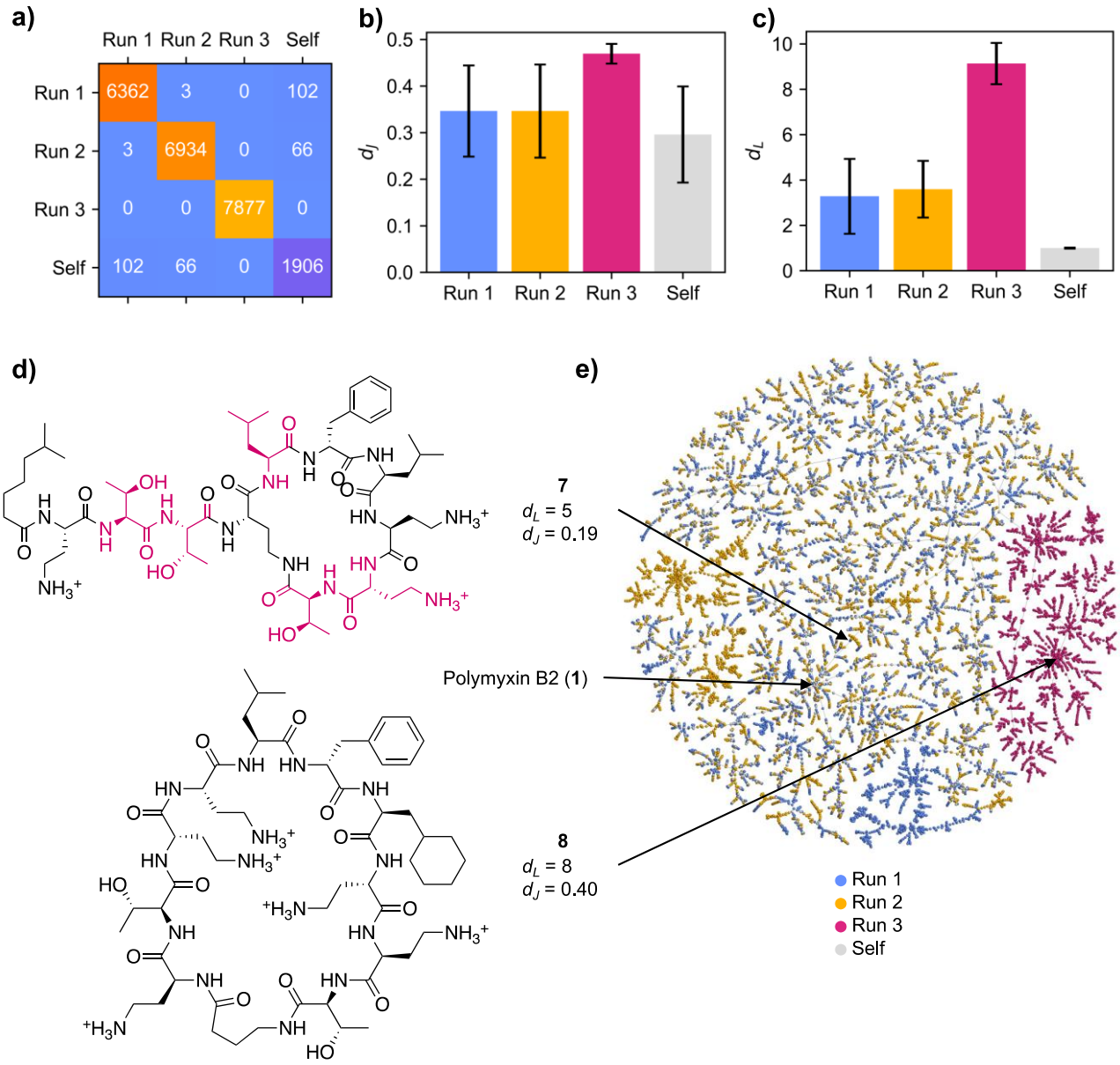


Figure 16. Analysis of polymyxin B2 runs starting from 50 random linear sequences (Run 1-3) or from polymyxin B2 without stopping condition (Self). a) Heatmap indicating the number of generated compounds with MAP4C $d_J < 0.5$ to polymyxin B2 for each trajectory, along with the number of overlapping compounds. b) Bar plot showing the mean and standard deviation of the d_J calculated using MAP4C fingerprints for generated compounds with $d_J < 0.5$ to polymyxin B2. c) Bar plot showing the mean and standard deviation of the Levenshtein distance (d_L; proxy for number of mutations) to polymyxin B2 for generated compounds with $d_J < 0.5$ to polymyxin B2. d) Structure of a selected polymyxin B2 analog featuring a high d_L and low d_J (7) and the closest analog generated in the failed run (8). e) TMAP displaying the generated compounds in a 2D space. Interactive TMAP: https://tm.gdb.tools/map4/10E60/polymyxin_randself_tmap.html.

5.3.3 Traversing chemical space to find median molecules

We next tested whether PDGA might be used to generate traversal trajectories in chemical space, starting from molecule A to reach a target molecule B, potentially travelling by a region of chemical

space containing median molecules, a goal realized by small molecule generation algorithms,[232,233] but not demonstrated for the case of peptides or peptide-like oligomers. PDGA was indeed able to generate such traversal trajectories between pairs of linear or cyclic peptides as illustrated with the pair of cyclic peptide natural products polymyxin B2 (**1**) and gramicidin S (**2**), the peptide/peptoid pair EB9 (**3**) and oncocin (**4**) and the pairs of linear 30-mers cathelicidin BF (**5**) and circulin D (**6**). Although reaching their targets, these trajectories rapidly diverged from the starting molecules and generated mostly close analogs to the target, without spending significant time at intermediate similarities (blue and red points in **Figure 17a** and **Figure C4**).

To obtain median molecules between A and B, we ran PDGA with a modified fitness function minimizing the sum of three terms, namely the Jaccard distances to A and B and their absolute difference. This fitness function guided the algorithm to produce molecules with the smallest possible but equal distance to A and B. Indeed, the population of molecules generated using this modified fitness function were close to the diagonal of the 2D-jaccard distance plot (yellow points in **Figure 17a** and **Figure C4**). A TMAP analysis of the set of molecules generated for the Polymyxin B2 (**1**) to gramicidin S (**2**) trajectories showed that each trajectory generated structurally distinct classes of molecules corresponding to different areas of the chemical space around these molecules, with interesting hybrid molecules such as **9** and **10** combining features from both compounds (**Figure 17b/c**).

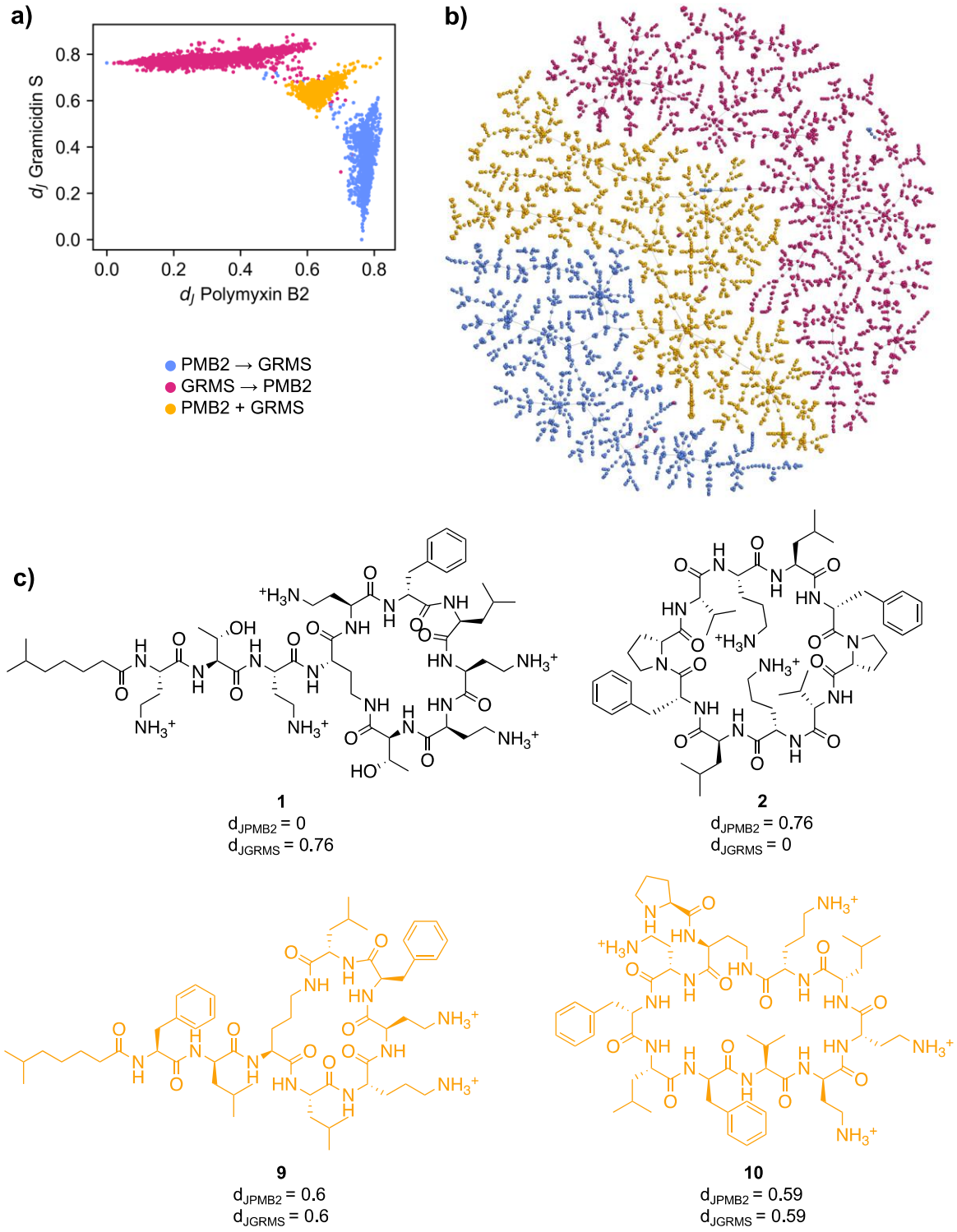


Figure 17. Visualization of traversal trajectories and median molecules between polymyxin B2 and gramicidin S. a) Jaccard distance of molecules selected from the different trajectories towards polymyxin B2 and gramicidin S. The trajectory from polymyxin B2 to gramicidin S is displayed in blue, the reverse trajectory is displayed in red, and the combined structure trajectory is displayed in yellow. b) MAP4C TMAP of selected molecules colored by their trajectory of origin. The trajectories populate separate chemical subspaces. c) Structures of the two queries polymyxin B2 and gramicidin S and two selected molecules from the median trajectory (yellow). Interactive TMAP: https://tm.gdb.tools/map4/10E60/polymyxin_gramicidin_tmap.html.

5.3.4 Traveling towards non-peptide molecules

We next used PDGA to identify analogs of targets not obtainable for the 100 selected building blocks, described here as "non-peptide", by minimizing the distance to target and stopping after 10,000 iterations. We tested this approach for diverse macrocycles containing building blocks and linkages not available in our library (**11-17**, **Figure C5**). For these non-peptide targets, driving PDGA with the shape and pharmacophore fingerprint MXFP delivered somewhat more convincing results than with MAP4C.

Specifically, the molecules generated using the MXFP fitness function matched the overall shape of the target molecules better than those generated using the MAP4C fitness function (**Figure 18** and **Figure C6**). For instance, in the case of cyclosporin (**11**), which contains several N-methylated amide bonds contributing to its membrane permeability,[234,235] and for valinomycin (**13**), where half of the linkages are ester instead of amide bonds, MAP4C generated macrocycles preserved more standard amide bonds, while those generated by MXFP guided PDGA to use the peptoid units available in our set of 100 building blocks, in order to mask the amide H-bond donor group. Furthermore, MAP4C sometimes selected acyclic analogs as best fits due to its emphasis on substructures, while MXFP always selected macrocycles matching the overall shape and polarity of the target molecule.

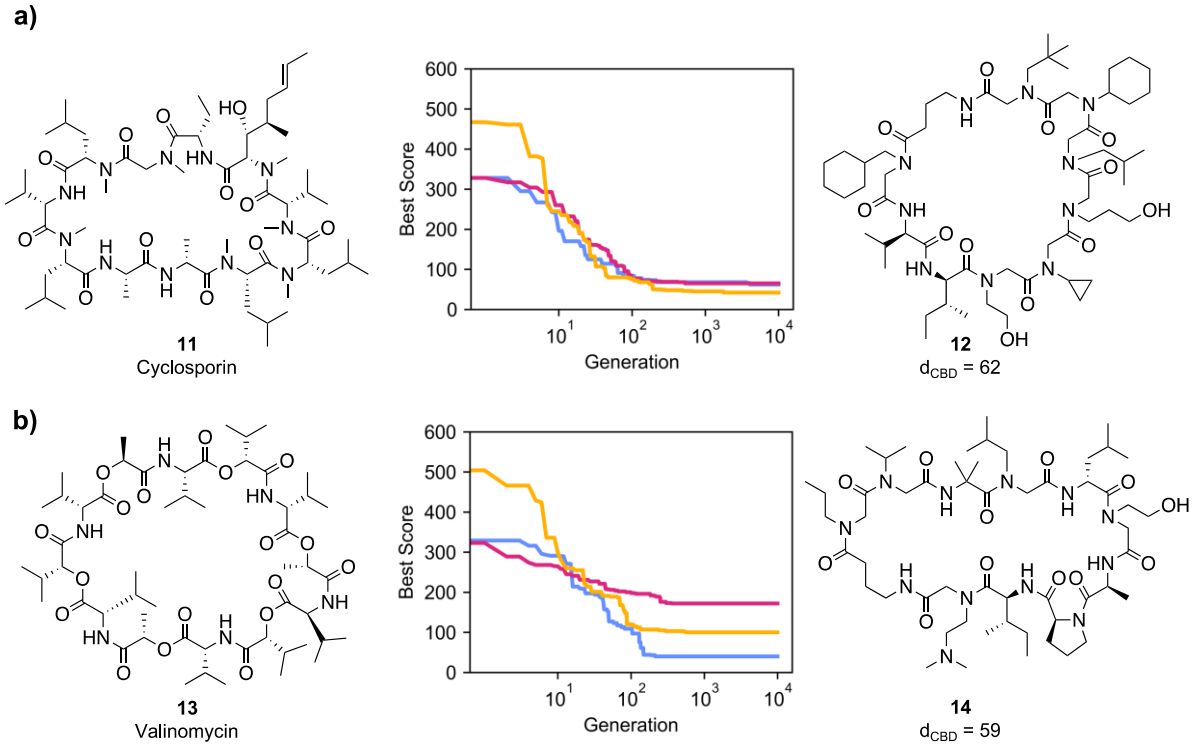


Figure 18. Non-peptide macrocycles, the overall best score throughout the trajectories and the corresponding best scoring MXFP analog from three combined runs for a) cyclosporin and b) valinomycin. The MXFP d_{CBD} is reported for each analog. See also Figure C6 for further details.

5.4 Conclusion

In the conversations around chemical space, 1E+60 has established itself as a symbolic and fascinating boundary. Here we explicitly created a virtual library of 1E+60 molecules by combining 100 peptide and peptoid buildings blocks to form up to 30-mer linear or cyclic oligomers, all potentially accessible by standard solid-phase synthesis. We demonstrated LBVS of this 1E+60 chemical space using a simple genetic algorithm, which succeeded in identifying virtual hits, defined either as analogs of specific molecules or as median molecules, by surveying only a few thousand sequences. It should be noted that, like in many journeys, the value of the chemical space journey using PDGA lies not in reaching the target but in the journey itself, here by encountering interesting molecules which would be otherwise difficult to design. Whether these molecules might translate into useful bioactives requires experimental evaluation of specific series. Additional studies along these lines are ongoing in our team.

5.5 Code availability

The code used for the analysis and plots study is available at https://github.com/reymond-group/10E60. The raw results files can be retrieved at https://zenodo.org/records/11396287.

5.6 Author Contribution Statement

MO designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper. Both authors read and approved the final manuscript.

5.7 Acknowledgements

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076).

6 Can large language models predict antimicrobial peptide activity and toxicity?

This chapter is based on a scientific article previously published in *RSC Medicinal Chemistry*. The article is reproduced here under the terms of the Creative Commons Attribution License (CC BY 3.0):

Orsi, M.; Reymond, J.-L. Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity? *RSC Med. Chem.* **2024**, *15* (6), 2030-2036.
https://doi.org/10.1039/D4MD00159A

6.1 Abstract

Antimicrobial peptides (AMPs) are naturally occurring or designed peptides up to a few tens of amino acids which may help address the antimicrobial resistance crisis. However, their clinical development is limited by toxicity to human cells, a parameter which is very difficult to control. Given the similarity between peptide sequences and words, large language models (LLMs) might be able to predict AMP activity and toxicity. To test this hypothesis, we fine-tuned LLMs using data from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP). GPT-3 performed well but not reproducibly for activity prediction and hemolysis, taken as a proxy for toxicity. The later GPT-3.5 performed more poorly and was surpassed by recurrent neural networks (RNN) trained on sequence-activity data or support vector machines (SVM) trained on MAP4C molecular fingerprint-activity data. These simpler models are therefore recommended, although the rapid evolution of LLMs warrants future re-evaluation of their prediction abilities.

6.2 Introduction

Antimicrobial peptides (AMPs) have gained significant attention in the field of drug discovery due to their potential therapeutic applications in the fight against antimicrobial resistance.[251-253] However,

the vast number of possible peptide sequences and their complex structure-activity relationship landscape mean that it is difficult to rationally design peptides with the desired biological activity, in particular tuning their activity versus toxicity to human cells, which is often measured as hemolysis of human red blood cells.[239,240]

To address this issue, several machine-learning models have been developed for the *de novo* design of antimicrobial peptides.[65,66,104-107,256-265] Because property prediction from a peptide sequence can be framed as a natural language processing problem, many of these models use architectures specifically designed for language processing tasks. [68,251,252] Furthermore, the emergence of large language models (LLMs), such as OpenAI's GPT models,[253] has opened new possibilities for leveraging powerful language processing capabilities in drug discovery applications. Recent attempts by Jablonka *et al*. to explore the capabilities of GPT-3 for predicting properties of small molecules in various applications have shown that GPT-3 was able to perform comparably or even outperform conventional statistical models, particularly in the low data regime.[254] There also have been successful efforts into augmenting LLM capabilities to tackle tasks related to small molecule chemistry in the areas of organic synthesis, drug discovery, and materials design.[72,270-272] Hereby, the models mainly orchestrate a set of tools to solve chemistry tasks starting from a natural language prompt.[74,258,259] However, to the best of our knowledge LLMs have not been implemented to predict the bioactivity of peptides yet.

In this study, we aimed to compare GPT models fine-tuned on antimicrobial peptide sequence data with models that have been previously used to predict antimicrobial activity and hemolysis of peptide sequences.[66,107] Alongside evaluating the performance of the fine-tuned GPT models, we also seek to explore the advantages and disadvantages they offer in terms of time and cost effectiveness. Furthermore, we compare the performance of models trained on amino acid sequences to a support-vector machine (SVM) trained on the MAP4C fingerprint.[42]

6.3 Methods

6.3.1 Datasets

The datasets used in this study were peptide sequences with annotated antimicrobial and hemolytic activity collected from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP).[66,260] Sequences exhibiting an activity measure below 10 mM, equivalent to 10,000 nM or 32 mg mL-1, against at least one of selected target organisms *P. aeruginosa*, *A. baumannii*, or *S. aureus* were categorized as active. Conversely, sequences with activity measures exceeding 10 mM, 10,000 nM, or 32 mg mL-1 against all of these targets were categorized as inactive. When available, activity against human erythrocytes was utilized to classify sequences as either hemolytic or non-hemolytic. Concentrations were standardized to mM, and sequences causing less than 20% hemolysis at concentrations equal to or above 50 mM were categorized as non-hemolytic and flagged accordingly. Sequences inducing more than 20% hemolysis were classified as hemolytic, irrespective of concentration. The dataset used for the classification tasks contained 9,548 (7,160 training / 2,388 validation) sequences with annotated antimicrobial activity, of which 2,262 (1,723 training / 539 validation) sequences had additional hemolytic activity annotations. To test models in low data regimes, we randomly selected subsets from the original training sets, representing approximately 20% and 2% of the original activity set, and approximately 10% of the original hemolysis set. All datasets are further described in **Table 4**.To ensure consistency, we maintained the same training and test split for all initial evaluations. For the detailed study, we used the same 5-fold cross-validation sets.

Table 4. Sizes and composition of the datasets used in the present study. Datasets are available at https://github.com/reymond-group/LLM_classifier.

Name	Size	# Positive Class	# Negative Class
Activity Training	7,160	3,580	3,580
Activity Training 20%	1,400	701	699
Activity Training 2%	140	74	66
Activity Validation	2,388	1,194	1,194
Hemolysis Training	1,723	717	1,006
Hemolysis Training 10%	35	65	105
Hemolysis Validation	539	226	313

6.3.2 Models

As reference models, we used our previously reported Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Recurrent Neural Network (RNN) classifiers trained on the same data.[66] We furthermore trained two additional SVM models on alternative representations of peptide sequences: one utilizing the MAP4C fingerprint[42] with a custom Jaccard kernel, and another using predicted fraction of helical residues and hydrophobic moment with a linear kernel. Fraction of helical residues were predicted using SPIDER3.[261] Hydrophobic moment was computed using the method of Eisenberg *et al.*[262]

To explore the potential of GPT-3 models for antimicrobial and hemolytic activity classification, we performed fine-tuning of the Ada, Babbage, and Curie models which were accessible through the OpenAI API (v0.28.0, accessed between 25.05.2023 and 01.06.2023). The fine-tuning process involved training each model using the full, 20% and 2% sets for activity classification and the full and 10% set for the hemolysis classification. In the later evaluation with the more advanced LLM GPT-3.5 Turbo, fine-tuning was also performed via OpenAI's Python API (v1.11.1), following the provided guidelines, but we restricted ourselves to the full model. The utilized fine-tuning datasets contained a system role ("predicting antimicrobial activity/hemolysis from an amino acid sequence"), a user message (peptide sequence formatted as "SEQUENCE ->"), and a system message ("0" for negative labels and "1" for positive labels).

6.3.3 Metrics

All models were evaluated using five commonly accepted performance metrics: ROC AUC, Accuracy, Precision, Recall and F1. Metrics were either calculated using the scikit-learn (v1.4.0) Python (v3.12.1) package (reference models and GPT-3.5) or directly obtained from the OpenAI platform after fine-tuning was completed (for all GPT-3 models).

ROC AUC (Receiver Operating Characteristic Area Under the Curve: The ROC AUC measures the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (Sensitivity) against the False Positive Rate. A higher ROC AUC value (ranging from 0 to 1) indicates better discrimination and predictive performance of the model.

Accuracy: Accuracy measures the overall correctness of the model's predictions, calculating the ratio of correctly classified instances to the total number of instances. It provides a general understanding of the model's performance but can be misleading in imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Precision: Precision measures the proportion of true positives out of all predicted positives. It focuses on the model's ability to avoid false positives.

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

Recall: Recall measures the proportion of true positives out of all actual positives. It represents the model's ability to identify positive instances accurately.

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

F1 score: F1 is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

6.4 Results and Discussion

6.4.1 Model screening

Starting from the DBAASP dataset of 9,548 peptide sequences annotated with antibacterial activity and 2,262 peptide sequences annotated with hemolysis effect, we had previously evaluated NB, RF, SVM and RNN models, and found the latter to perform best for predicting both activity and hemolysis from sequence data.[66,107] For additional reference, we trained an SVM on the fraction of helical residues and the hydrophobic moment, two properties commonly known to correlate with antimicrobial activity, as well as another SVM on MAP4C, a molecular fingerprint that can reliably encode large molecules such as natural products and peptides including their chirality,[42] a parameter which we considered important since our data listed sequences containing both L- and D-amino acids.

Aiming to test how LLMs perform in predicting antimicrobial activity and hemolysis, we first fine-tuned and evaluated GPT-3 Ada, Babbage, and Curie models (**Table D1**). As discussed in our preprint, these models performed slightly better than the reference models and even provided good performances when trained in low data regime (20% and 2% of full data). However, these models were later deprecated by OpenAI, and their performance cannot be reproduced. We therefore discuss herein only the results obtained with the more recent GPT-3.5 model, in comparison with the reference models.

For both, prediction of antimicrobial activity and prediction of hemolysis, the top-performing models were the MAP4C SVM and the RNN model trained on sequence data, the latter being the best performer in our original work (**Table 5**).[66] The performances for both models were in a similar range, although the RNN displayed a notably higher ROC-AUC in both tasks. GPT-3.5 displayed the highest recall performance among the activity models, indicative of the model's tendency to overly favour positive predictions, potentially leading to increased false positive predictions. On the other hand, the features SVM trained only on helicity and hydrophobic moment did not perform significantly above background and was later used as a negative control model.

Table 5. Performance metrics of all models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold. NB: Naïve Bayes, RF: Random Forest, SVM: Support Vector Machine, RNN: Recurrent Neural Network, MAP4C: Chiral MinHashed Atom-Pair Fingerprint of Diameter 4, GPT: Generative Pre-Trained Transformer.

Model	ROC AUC	Accuracy	Precision	Recall	F1
NB act.	0.55	0.55	0.59	0.32	0.42
RF act.	0.81	0.71	0.7	0.75	0.73
SVM act.	0.75	0.68	0.68	0.68	0.68
RNN act.	**0.84**	0.76	0.74	0.8	0.77
Features SVM act.	0.65	0.65	0.66	0.62	0.64
MAP4C SVM act.	0.8	**0.8**	**0.79**	0.83	**0.8**
GPT-3.5 Turbo act.	0.68	0.68	0.62	**0.93**	0.75
NB hem.	0.58	0.56	0.48	0.76	0.59
RF hem.	0.8	0.77	**0.81**	0.6	0.69
SVM hem.	0.69	0.73	0.72	0.58	0.65
RNN hem.	**0.87**	0.76	0.7	0.76	0.73
Features SVM hem.	0.62	0.63	0.57	0.5	0.54
MAP4C SVM hem.	0.83	**0.83**	0.76	**0.85**	**0.8**
GPT-3.5 Turbo hem.	0.65	0.69	0.72	0.43	0.54

6.4.2 Model comparison

Following the initial model screening, we aimed to validate our findings through a more robust approach: a 5-fold cross-validation involving GPT-3.5, the MAP4C SVM, the RNN, and finally the features SVM as negative control. For this purpose, we generated five data splits and conducted predictions anew.

The results, depicted in **Figure 19a** for antimicrobial activity prediction and **Figure 19b** for hemolysis prediction, confirmed our earlier observations (performances in **Table D2)**. Notably, the RNN performances were higher than those observed in the screening experiment and were clearly above those of GTP-3.5. Furthermore, both the RNN and MAP4C SVM demonstrated comparable

performances, indicating the validity of both approaches in predicting antimicrobial activity and hemolysis. The finding that simpler machine learning architectures, like SVM, can rival the performance of more complex RNNs in predicting antimicrobial activity and hemolysis is particularly interesting. A comparison with models trained on similar datasets, which achieve similar performances as reported in this study, further reinforces the consistency of our findings.[263-265]

This raises questions about the importance of model architecture versus foundational elements such as data quality and feature engineering. It suggests that a balanced approach, prioritizing optimization of these foundational components, could prove more beneficial than focusing solely on model complexity.

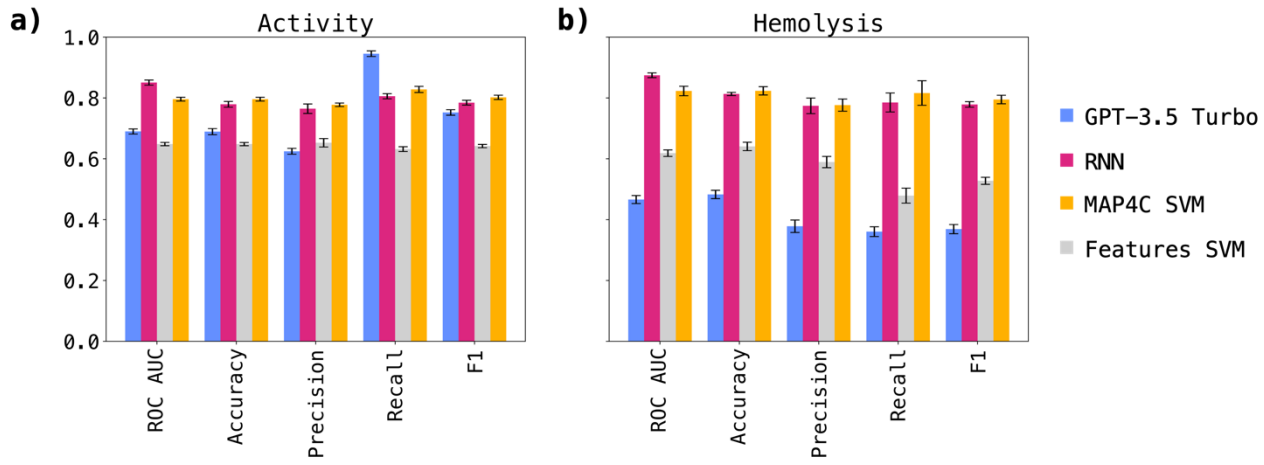


Figure 19. Results of the 5-fold cross-validation study aimed at validating MAP4C SVM, Features SVM, RNN, and GPT-3.5 turbo performance for a) antimicrobial activity and b) hemolysis predictions. The mean performance across the 5 cross-validations for each metric is shown as a bar, the standard deviation is displayed with an error bar. The results confirmed earlier observations but showed notably higher performances for the RNN compared to the one-shot screening experiment. Both the RNN and MAP4C SVM demonstrated comparable performances.

6.4.3 Data visualization

The high performance achieved by the SVM trained on the MAP4C fingerprint suggested that the nearest neighbour relationships in the MAP4C feature space could be sufficient to distinguish active from inactive and hemolytic from non-hemolytic peptide sequences. In our previous work, we observed that the MAP4 fingerprint[40] correctly clustered natural products, taken from the COCONUT database,[173] according to their organism of origin.[263,115] In analogy to our previous work, we were

curious to see whether a spatial separation of actives/inactives and hemolytic/non-hemolytic sequences can be obtained from encoding with MAP4C, the chiral version of MAP4, possibly explaining the good performance of the MAP4C SVM model. For this, we reduced the 2048-dimensional feature space of MAP4C to 2D using the dimensionality reduction method TMAP,[41] and used the obtained visualization to display a set of molecular properties.

First, we wanted to confirm that the TMAP visualization aligns with intuitive distributions of structural features relevant for peptides. For that, we coloured the data points based on their heavy atom count (HAC), an indicator of molecular size, and fraction of carbon atoms (fraction C), a simple proxy for the hydrophobicity of a peptide sequence. The TMAP revealed visible clusters for both, HAC (**Figure 20a**) and fraction C (**Figure 20b**), indicating that the reduced MAP4C features can reliably represent simple molecular descriptors in the underlying chemical space.

Following this first observation, we wanted to test if we can detect clusters within TMAP visualizations of more complex physicochemical properties, such as the predicted fraction of helical residues (**Figure 20c**) and the hydrophobic moment (**Figure 20d**). In both cases, we could not detect large homogenous clusters as was the case for HAC and fraction C. However, the data formed a large number of small local clusters, indicating that the nearest neighbour relationships in the MAP4C feature space can possibly be used to distinguish sequences with high helicity/hydrophobicity opposed to sequences with low helicity/hydrophobicity.

Finally, we analysed the distribution of active versus inactive (**Figure 20e**) and hemolytic versus non-hemolytic (**Figure 20f**) sequences in the MAP4C chemical space. Similarly to the visualizations of predicted fraction of helical residues and hydrophobic moment, active and inactive or hemolytic and non-hemolytic sequences are spatially separated in a large number of small, local clusters. This finding is particularly interesting as it suggests that nearest neighbour relationships in the MAP4C feature space are sufficient to separate peptide sequences based on their antimicrobial activity and hemolysis. It further provides an explanation to the good performance obtained with the MAP4C SVM, which can leverage the nearest neighbour relationships stored in the MAP4C fingerprint feature space when provided with a custom Jaccard kernel function.

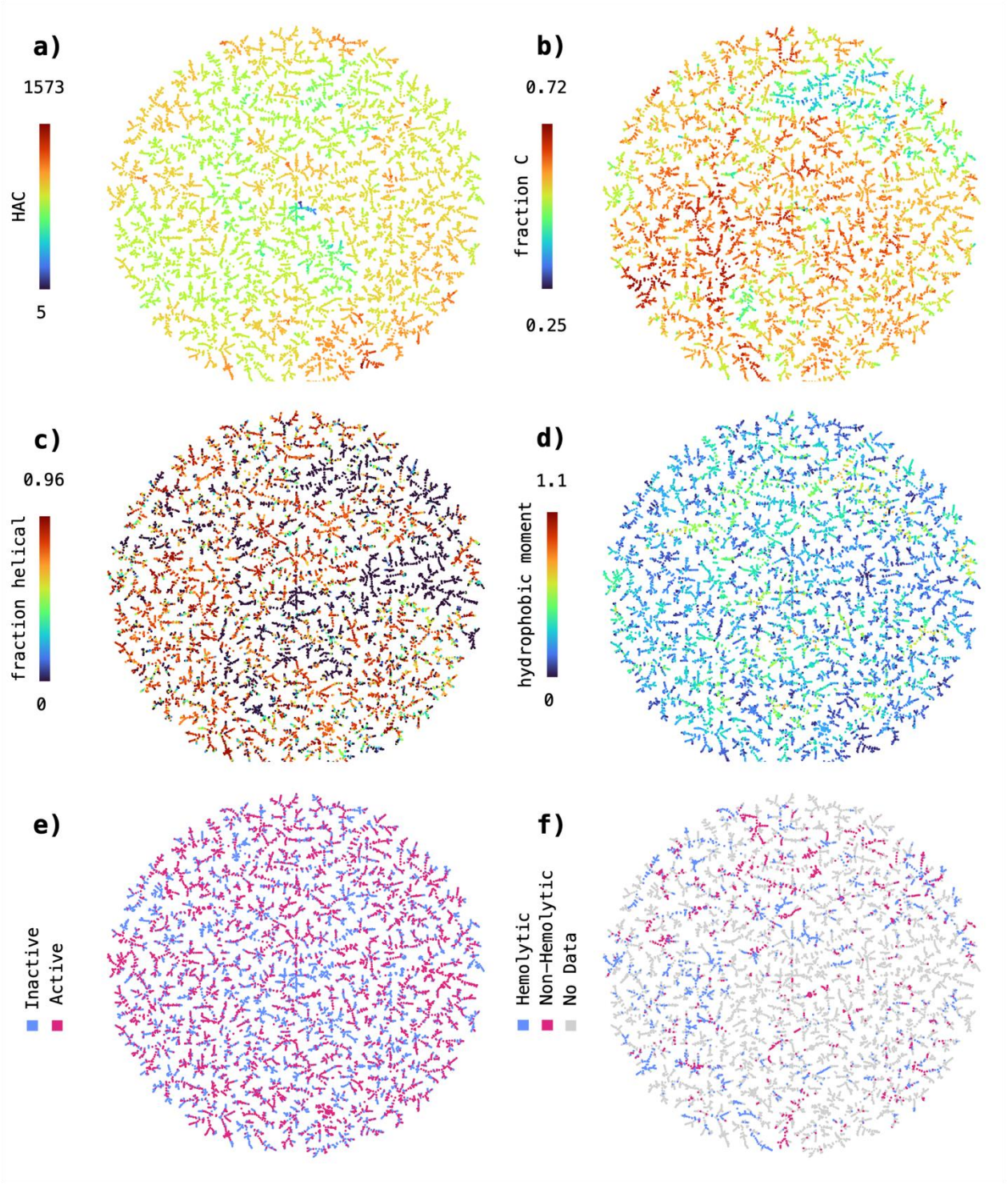


Figure 20. Chemical space covered by the 9,548 peptide sequences with annotated antimicrobial activity extracted from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP). The sequences are encoded using the MAP4C fingerprint and the resulting 2048-dimensional space reduced to 2D using TMAP. The sequences in the 2D TMAP were colored based on a) heavy atom count, b) fraction of carbon atoms, c) predicted fraction of helical residues, d) hydrophobic moment, e) annotated antimicrobial activity and f) annotated hemolysis.

6.5 Conclusion

In the present study we investigated the potential of LLMs as predictive tools for antimicrobial activity and hemolysis of peptide sequences. We assessed that fine-tuning GPT models in cloud is a relatively easy and fast process as access through the API eliminates the need to buy expensive hardware and requires little technical expertise. Duration of fine-tuning was short, and the associated costs were low (**Table D3**). In contrast to cloud-based fine-tuning, local model training involves setting up and maintaining hardware, which can be costly and require technical expertise. While less complex models like RNNs and SVMs have lower hardware requirements, training larger models such as LLMs locally can pose challenges in terms of scalability, as one can rapidly face limitations in terms of hardware capacity and maintenance costs.

However, the lack of control over the training environment in cloud-based approaches raises concerns regarding reproducibility of scientific results. In the course of this study, we had originally fine-tuned GPT-3 models Ada, Babbage and Curie. These models performed slightly better than the reference models, even achieving good performances in low data regimes. Unfortunately, these models were later deprecated by OpenAI and their performance cannot be reproduced. When fine-tuning a newer iteration of GPT-3 (GPT-3.5 Turbo), we observed a significant decrease in performance for the same task. We attribute the drop in performance to the increasing optimization of LLMs for conversational interactions, which may negatively impact their effectiveness in out-of-scope predictive tasks. These findings highlight the potential risk of how not controlling one's own models can compromise the reproducibility and reliability of scientific results.

The aforementioned findings suggest a diminishing suitability of chat oriented LLMs for classification tasks over time, a function beyond their intended design. This observation specifically applies to LLMs tailored for conversational or human interaction purposes, rather than specialized LLMs trained on domain-specific data. Unfortunately, the latter do not provide the ease of access and usability that GPT models do. Consequently, we expect that LLMs will increasingly be employed in

human interaction settings, facilitating the integration of various chemical tools through natural language interfaces as is being pioneered by Bran[258] and Boiko *et al.*[259]

Finally, we could demonstrate in the present study that classical machine learning techniques, such as SVMs trained on MAP4C fingerprint encodings, can achieve state-of-the-art performance in the prediction of antimicrobial activity and hemolysis. This finding is especially interesting, as it showcases that good performance can be achieved by less complex models, putting the emphasis on data quality rather than model complexity.

6.6 Code availability

The source codes and datasets used for this study are available at https://github.com/reymond-group/LLM_classifier.

6.7 Author Contribution Statement

MO designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper. Both authors read and approved the final manuscript.

6.8 Acknowledgements

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076). MO thanks Sacha Javor for the helpful discussion and comments.

7 Assigning the stereochemistry of natural products by machine learning

Orsi, M.; Reymond, J.-L. Assigning the Stereochemistry of Natural Products by Machine Learning. September 23, **2024**. https://doi.org/10.26434/chemrxiv-2024-zz9pw.

Abstract

Nature has settled for L-chirality for proteinogenic amino acids and D-chirality for the carbohydrate backbone of nucleotides. Here we asked the question whether stereochemical patterns might also exist among natural products (NPs) such that their stereochemistry could be assigned automatically. Indeed, we report that a language model can be trained to assign the stereochemistry of NPs using the open access NP database COCONUT. In detail, our language model, called **NPstereo**, translates an NP structure written as absolute SMILES into the corresponding isomeric SMILES notation containing stereochemical information with 80.1% per-stereocenter accuracy for full assignments and 86.3% per-stereocenter accuracy for partial assignments across various NP classes including secondary metabolites such as alkaloids, polyketides, lipids and terpenes. **NPstereo** might be useful to assign or correct the stereochemistry of newly discovered NPs.

7.1 Introduction

Since the identification of carbon containing molecules as signature constituent of living matter on our planet (*vis vitalis*), deciphering the structure and function of natural products (NPs) has guided the development of organic chemistry and remains an essential source of inspiration for the development of new medicines.[282-293] However, while the atom connectivity of NPs can be assigned by a variety of methods ranging from degradation chemistry to Mass Spectrometry and NMR spectroscopy, as of today determining the configuration of individual stereocenters in NPs (3D-structure) remains challenging and requires techniques such as X-ray crystallography and chiroptical spectroscopy often combined with derivatization,[294-299] and sometimes total synthesis to confirm or correct the initial stereochemical assignment.[300-308]

Considering that nature has settled for only L-chirality in proteinogenic amino acids and D-chirality in the carbohydrate backbone of nucleotides, we asked the question whether similar regularities might be hidden in NP stereochemistry that might allow this information to be machine learned. To the best of our knowledge, this question has not been addressed despite many studies using machine learning to classify NPs and their relation to other molecular classes.[114,156,280,309-313] We set out to test if a transformer model, a type of neural network initially developed for language translation,[68] found to perform well for a variety of chemistry related tasks,[296] including the prediction of stereoselective reactions in forward and retrosynthesis direction,[297–299] might be able to learn NP stereochemical assignments.

As detailed below, we found that the stereochemistry of NPs can indeed be assigned by a transformer model trained on inserting missing stereochemical labels for chiral centers and *Z/E* double bonds into a SMILES (Simplified Molecular Input Line Entry System)[171,172] string representation of the molecular structure with entirely missing or partially assigned stereocenters. We trained our model using 63,998 NPs with fully assigned stereochemistry and a literature reference, which we collected from the open access database COCONUT (COlleCtion of Open Natural ProdUcTs), a database which combines several NP databases into a single source.[173] Our transformer

model, called **NPstereo**, assigns NP stereochemistry with 80.1% per-stereocenter accuracy for full assignments and 86.3% per-stereocenter accuracy for partial assignment across various NP classes.

7.2 Results and Discussion

7.2.1 Dataset analysis

All NPs associated with at least one associated literature reference were extracted from the COCONUT database, which provided 116,403 NPs written as canonical isomeric SMILES, including stereochemical information for tetrahedral centers (@ or @@) and stereogenic double bonds (/C=C/ or /C=C\). Visualizing similarity relationships between NPs as measured by the MAP4C molecular fingerprint,[42] using the dimensionality reduction method TMAP,[39,41] provided a layout illustrating the different structural classes (polyketides, benzenoids, nucleosides, alkaloids, lignans, peptides, lipids & terpenes, and glycosides, **Figure 21a**). The structural classes differed by the number of stereocenters per molecule, which was relatively high for lipids & terpenes, peptides, and glycosides, and much lower for benzenoids and nucleosides (**Figure 21b**). Some of the 116,403 NPs corresponded to different stereoisomeric forms of the same 2D-structure, including structures with incomplete of fully missing stereochemical assignment, such that the set only contained 98,108 different 2D-structures without stereochemical assignment, written as absolute SMILES. Note that NPs with incomplete stereochemical assignment were evenly distributed among the different NP structural classes, as evidenced by color-coding the TMAP (**Figure E1**).

The subset of NPs with fully assigned stereochemistry featured 73,130 different isomeric SMILES corresponding to 63,998 different absolute SMILES after removal of stereochemical labels. Among these, 12,095 absolute SMILES (18.9%) did not contain any stereocenter, 9,254 (14.5%) contained 11 or more stereocenters, and the remaining 42,649 (66.6%) were approximately evenly distributed in groups containing between one and ten stereocenters. Most of these absolute SMILES corresponded to a single stereoisomer (i.e. a single isomeric SMILES, 56,676, 88.6%), a small group

to two stereoisomers (6,100, 9.5%), and a very small fraction (1,212, 1.9%) to three or more stereoisomers (**Figure 21c**).

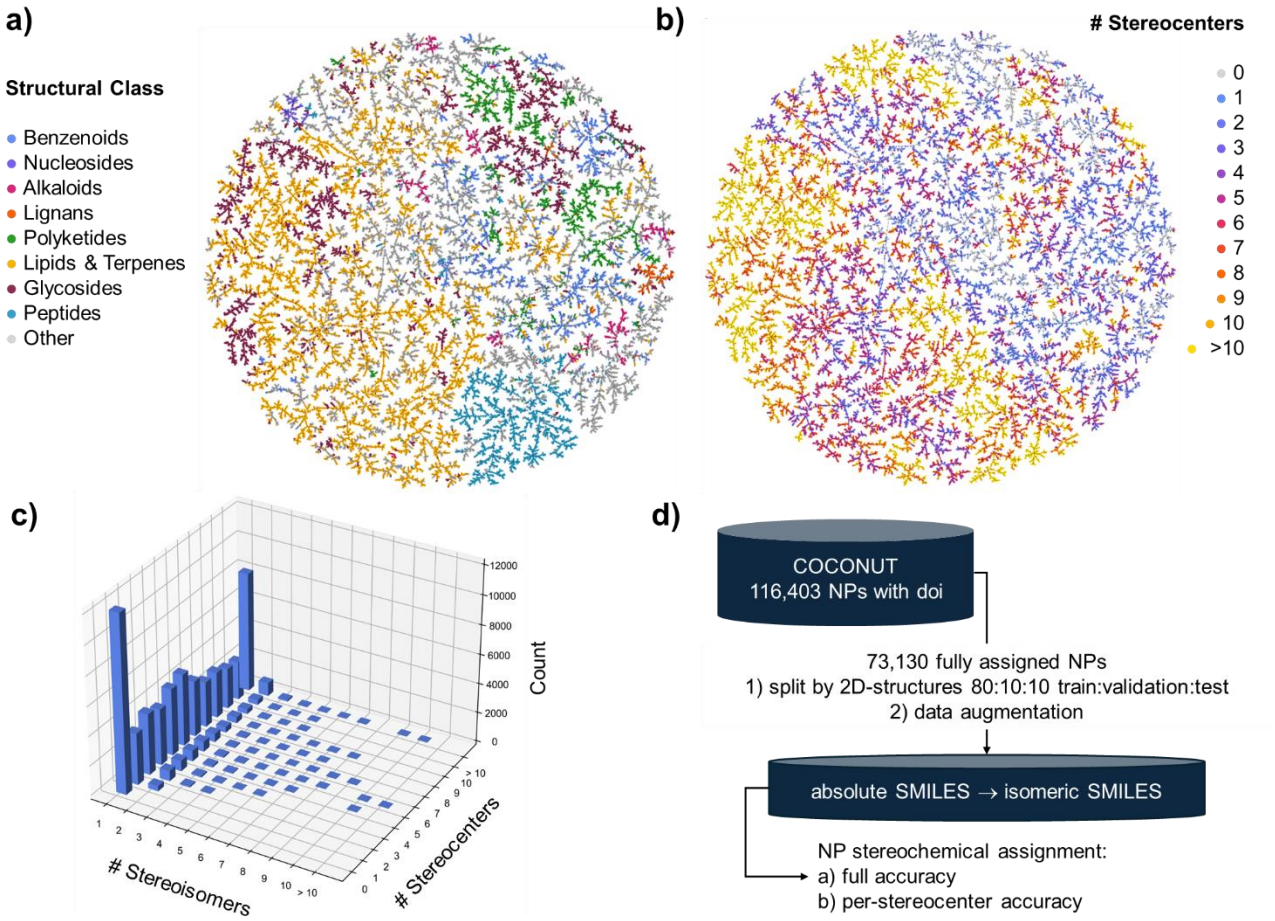


Figure 21. Dataset analysis and model training strategy. (a-b) MAP4C TMAPs of 116,403 unique compounds with an associated DOI, extracted from the COCONUT database. The TMAP visualizations are colored according to a) NP structural classes, as defined in the COCONUT database "chemical_super_class" field, with peptides and glycosides further refined using SMARTS substructure searches; and b) the number of stereocenters in each structure. An interactive version of the TMAP plot can be accessed at: https://tm.gdb.tools/map4/NPstereo/. (c) 3D bar plot showing the number of molecules (vertical axis) as a function of the number of total stereocenters (depth axis) and the number of stereoisomers (horizontal axis) for 63,988 structures with fully assigned stereocenters. (d) Workflow for data selection, model training and evaluation.

7.2.2 Model design and training

To test if NP stereochemical assignment could be machine learned, we set out to train transformer models to translate a source absolute SMILES, describing the unassigned 2D-structure of an NP, into the corresponding target isomeric SMILES containing stereochemical labels. For model training, we

used the 73,130 NPs with a literature reference and a fully assigned stereochemistry extracted from COCONUT. We split the data into training, validation and test sets at the level of 2D-structures (63,988 canonical absolute SMILES, lacking stereochemical labels) with an 80:10:10 ratio, and optionally considered data augmentation schemes compensating for the relatively small dataset size to train and evaluate various models in terms of full assignment accuracy and per-stereocenter assignment accuracy (**Figure 21d**).

To train our first transformer model **C1**, we considered each canonical isomeric SMILES in the training and validation splits and generated a corresponding absolute SMILES by removing all stereochemical labels, which resulted in an absolute SMILES with the same order of characters as the canonical isomeric SMILES. These absolute SMILES were used as source strings and associated with the corresponding canonical isomeric SMILES as target strings, resulting in training and validation datasets for model **C1** (see methods for details). In this manner, model **C1** would be trained to convert each absolute SMILES into the corresponding canonical isomeric SMILES by inserting stereochemical labels without having to alter the order of characters in the SMILES.

Next, we enlarged the training and validation sets of **C1** using two possible data augmentation approaches. First, we used SMILES randomization,[300] a technique which generates a number of non-canonical forms of a canonical SMILES and is often used to enhance the performance of language models trained on SMILES.[301,302] We applied the procedure to the target lists of canonical isomeric SMILES at 2-, 5-, 10-, 20-, and 50-fold levels, and subsequently removed stereochemical labels from the resulting non-canonical isomeric SMILES to produce the corresponding absolute SMILES for the augmented source lists, which resulted in augmented datasets to train models **A2**, **A5**, **A10**, **A20** and **A50**. Second, we randomly removed stereochemical labels from each canonical isomeric SMILES of the target list in **C1** in up to five different versions for each number of removed label to produce an augmented source list of partially assigned isomeric SMILES. We then paired each of these partially assigned isomeric SMILES with their parent fully assigned canonical isomeric SMILES in the target list. This procedure resulted in a 25-fold augmentation of the data to train model **NPstereo**. Finally, we combined both data augmentation approaches by applying a 10-fold SMILES randomization to

the target lists of canonical isomeric SMILES of **C1**, and then randomly removing stereochemical labels from each randomized isomeric SMILES one at a time until none remained to produce partially assigned randomized SMILES for the source list. This procedure augmented the **C1** training and validation datasets by approximately 65-fold, resulting in training data for model **M65**.

In addition to these augmented datasets, we generated a negative control dataset **R1** by randomizing stereochemical labels in the target lists of canonical isomeric SMILES of model **C1** such that no pattern in stereochemistry should be recognizable. An additional negative control dataset **RP** was created by partial removal of stereochemical label from the target list of **R1** to augment the source list. All models described above were trained for approximately 9 hours to complete 100,000 steps (see methods for details).

7.2.3 Performance evaluation

We first tested the different models on writing fully assigned canonical isomeric SMILES from absolute SMILES, a task which corresponds to assigning the configuration of all stereocenters in an NP 2D-structure (**Table 6**, upper part, center: canonical full assignment test set). All models except the negative control model **R1** produced almost exclusively (>99%) valid SMILES, indicating reliable learning of the canonical SMILES syntax. In terms of prediction accuracy, the best performing model was **NPstereo** trained on canonical SMILES including partially assigned sources, which achieved 58.1% top-1 accuracy for full assignment and 80.1% top-1 accuracy per assigned stereocenter. The second-best model was **A50** with 56.3% top-1 accuracy for full assignment and 80.3% top-1 accuracy per assigned stereocenter. All other models performed worse but still significantly above the negative control models **R1** and **RP** trained with randomized stereochemical labels, which achieved 23-24%% top-1 accuracy for full assignment. This performance level reflected their ability to identify NPs lacking stereocenters (19% of the dataset), combined with the probability that a random stereochemical assignment can be correct (50% for the 6% NPs containing a single stereocenter). The inability of both negative control models to recognize stereochemical patterns was well reflected in their ~50% top-1 per-stereocenter accuracy, close to the random guess expectation.

When tested on non-canonical absolute SMILES, **NPstereo** only produced 83.4% valid SMILES and performed at the same levels as the negative control models **R1** and **RP** for full assignment and per-stereocenter accuracy (**Table 6,** upper part, right, non-canonical full assignment test set). This low performance indicated that **NPstereo**, which was trained with canonical SMILES, needed the canonical order of SMILES characters in the source to produce valid SMILES annotated with the correct NP stereochemistry. On the other hand, models **A2**-**A50** trained with both canonical and non-canonical SMILES performed similarly well on both SMILES types in terms of SMILES validity (98 - 99.3%), top-1 accuracy for full assignment (45.4 - 56.2%), and top-1 per-stereocenter accuracy (72.5 - 80%). The mixed model **M65** trained with canonical and non-canonical partially assigned source SMILES, performed similarly to model **A5** in terms of SMILES validity (99.2%), top-1 accuracy for full assignment (47.7%) and per-stereocenter top-1 accuracy (75.9%), indicating that the addition of partially assigned SMILES was not helpful for learning the stereochemistry of NP written in non-canonical SMILES format.

Models **NPstereo**, **M65** and the negative control **RP**, trained on translating partially assigned isomeric SMILES to the corresponding fully assigned SMILES, were additionally tested on adding missing stereochemical labels to partially assigned SMILES. This task is comparable to completing the stereochemical assignment of a partially assigned NP structure, which is often encountered in practice. When tested with partially assigned canonical SMILES, **NPstereo** was better than **M65** and performed slightly better than on the full assignment task in terms of top-1 overall accuracy (64.6%) and top-1 per-stereocenter accuracy (86.3%, **Table 6**, lower part, center, canonical partial assignment test set). The model however collapsed to the level of the negative control on all three measures when tested with non-canonical SMILES, indicating again the requirement for a canonical order of characters in the source SMILES for proper prediction. On the other hand, model **M65** performed quite well for assigning missing stereocenters to partially assigned non-canonical SMILES in terms of SMILES validity (99.1%), top-1 overall accuracy (54%) and top-1 per-stereocenter accuracy (82.1%, **Table 6,** lower part, right, non-canonical partial assignment test set).

Table 6. SMILES validity and performance metrics of models for NP stereochemistry assignment evaluated across different dataset augmentation strategies.

	Training Dataset (Train + Validation)		**Canonical Full Assignment Test Set[a)]**			**Non-Canonical Full Assignment Test Set[b)]**		
Model	SMILES type	Size	SMILES Validity[c)]	Full-Assignment Accuracy[d)] Top 1/ 2 / 3	Per-Stereocenter Accuracy[e)] Top 1 / 2 / 3	SMILES Validity[c)]	Full-Assignment Accuracy[d)] Top 1/ 2 / 3	Per-Stereocenter Accuracy[e)] Top 1 / 2 / 3
C1	Absolute → Canonical	58,571	99	56.3 / 67.4 / 71.4	78.7 / 86.9 / 89.2	76.7	22.5 / 27.3 / 30.5	41 / 49.7 / 53.9
A2	Absolute → Randomized (2x)	116,872	99.4	36.2 / 46.6 / 52.5	67.8 / 78.3 / 82.4	98	45.4 / 56.1 / 61.5	72.5 / 81.7 / 85.4
A5	(5x)	288,472	99.6	44.2 / 56.4 / 62.3	73.9 / 83.7 / 87.3	99.1	50.4 / 62.2 / 68.3	76.8 / 85.9 / 89.3
A10	(10x)	570,898	99.7	51.7 / 65 / 71.2	77.9 / 87.3 / 90.8	99.4	54.4 / 67.3 / 73.6	79.3 / 88.2 / 91.5
A20	(20x)	1,117,782	99.7	55.1 / 68.6 / 75.1	79.7 / 89 / 92.1	99.4	55.6 / 68.7 / 74.8	79.8 / 88.8 / 91.8
A50	(50x)	2,651,393	99.6	56.3 / 69.6 / 75.6	**80.3** / 89.1 / 92.3	99.3	**56.2** / 69.9 / 76.1	**80** / 89.1 / 92.3
NPstereo	Partially Assigned → Canonical	1,370,809	99.4	**58.4** / 69.9 / 75.4	80.1 / 88.6 / 91.4	83.4	23 / 28.7 / 32.3	45.4 / 55.3 / 59.8
M65	Partially Assigned → Canonical or Randomized	3,872,215	99.7	46.4 / 59.5 / 66.1	75.3 / 85.2 / 88.9	99.2	47.7 / 60.6 / 66.7	75.9 / 85.4 / 88.9
R1	Absolute → Canonical with randomized stereochemistry	58,571	92.1	23.1 / 28.9 / 31.9	49.9 / 59.2 / 63	97.5	23.3 / 30.1 / 33.5	50.9 / 62.7 / 66.8
RP	Partially Assigned → Canonical with randomized stereochemistry	1,757,098	99.1	24 / 30.5 / 34.6	52.6 / 64.8 / 69.6	80.9	19.6 / 24 / 27.5	39.2 / 48.5 / 52.7
			Canonical Partial Assignment Test Set[f)]			**Non-Canonical Partial Assignment Test Set[g)]**		
NPstereo	Partially Assigned → Canonical	1,370,809	99.5	**64.6** / 76.6 / 80.6	**86.3** / 92.6 / 94.3	78.1	14.2 / 22.3 / 26.5	46 / 56.6 / 61.4
M65	Partially Assigned → Canonical or Randomized	3,872,215	99.8	51.7 / 68.6 / 75.1	81.4 / 90.9 / 93.5	99.1	**54** / 70.5 / 76.5	**82.1** / 91.1 / 93.6
RP	Partially Assigned → Canonical with randomized stereochemistry	1,757,098	99.3	13.9 / 26.6 / 32.7	53.3 / 69.1 / 74.3	75.5	8.1 / 15.4 / 19.2	37.1 / 48.8 / 53.5

a) Test set consisting of absolute SMILES generated by removing all stereochemical labels from canonical isomeric SMILES. b) Test set consisting of absolute SMILES generated by removing all stereochemical labels from randomized isomeric SMILES. c) Percentage of valid SMILES generated by the model considering the top-3 outputs. d) Full-assignment accuracy represents the percentage of times the isomeric SMILES (canonical or non-canonical) of the target is produced by the model in the top-1, top-2 or top-3 outputs in response to the source absolute or partially assigned isomeric SMILES (canonical or non-canonical). e) The per-stereocenter accuracy is the highest percentage of correctly predicted stereocenters per molecule when analyzing the top-1, top-2 and top-3 isomeric SMILES outputs of the model. For NPs without stereocenter, the predicted SMILES is considered correct if it matches the target SMILES. The best top-1 value for each accuracy metric is highlighted in bold. See text and methods for details. f) Test set consisting of isomeric SMILES generating by removing part of the stereochemical labels from canonical isomeric SMILES. g) Test set generated by removing part of the stereochemical labels from randomized isomeric SMILES.

Analysis of model performance as function of the number of unassigned stereocenters per molecule provided further insights into model performance. For full stereochemical assignment on the canonical SMILES test set, the best models **NPstereo** and **A50** were closely matched in performance across all numbers of stereocenters per molecule (light green and dark blue curves in **Figure 22a**, left panel). However, these models were surpassed by **NPstereo** tested on the partial stereochemical assignment test set with up to nine unassigned stereocenters, indicating that, not unexpectedly, the availability of partial stereochemical information helped the model to assign the missing stereocenters (green curve in **Figure 22a**, left panel). Note that the higher performance of all models for NPs with five or less stereocenters partly reflected the contribution of a chance assignment to be correct, as indicated by the performance curve of the negative control models **R1**, **RP** and **RP** on partial assignment task, which matched the performance expected from chance assignment (light grey, grey, black and dashed black lines, **Figure 22a**, left panel). This analysis also showed that all models including the negative controls trained with randomized stereochemical labels performed perfectly with NPs lacking stereocenters, implying that they had learned not to add any stereochemical labels when no stereocenters were present.

For the non-canonical SMILES test set, the top-1 assignment accuracy as function of the number of unassigned stereocenters highlighted the performance collapse of models **C1** and **NPstereo** to random and negative control levels for all number of stereocenters in both the full and the partial stereochemical assignment tasks, an effect also apparent in Table 1 discussed above (red and light green lines, **Figure 22a**, right panel). In fact, models only trained on canonical SMILES (**C1**, **NPstereo**, and the negative controls **R1** and **RP**) performed even below chance levels and even partly failed to identify NPs without stereocenters for this non-canonical SMILES test set. On the other hand, model **A50** performed quite well for the full assignment and model **M65** for the partial assignment task on these non-canonical SMILES on which they had been trained (dark blue and dark yellow lines, **Figure 22a**, right panel). The dependence of model performance on the SMILES syntax (canonical or non-canonical) in training versus test sets showed that stereochemical assignment was learned according to the order of characters in the SMILES, which is very different in randomized

versus canonical SMILES. This effect also explained the need for a relatively large training set (50-fold augmentation) for a model to learn stereochemistry in the more diverse context of randomized SMILES.

Analysis of the top-1 per-stereocenter assignment accuracy as function of the number of stereocenters showed that, for the full assignment task on canonical SMILES (all models except negative controls, **Figure 22b**, left panel) and non-canonical SMILES (all models trained with randomized SMILES, **Figure 22b**, right panel), performance was lowest for NPs with a single chiral center, and gradually increased with additional stereocenters. These effects probably reflected the presence of regularities in NPs with high number of stereocenters such as the homochirality in most peptides and oligonucleotides and the limited stereochemical diversity of carbohydrates and steroids, because such regularities would be easier to learn for the different models in the full assignment task.

The same effect could explain the almost constant performance of the partial assignment task as the number of unassigned stereocenters increased for models **NPstereo** (green line in **Figure 22b**, left panel) and **M65** (dark yellow line in **Figure 22b**, both panels), because the difficulty to assign stereochemistry to NPs with a single stereocenter would be compensated by the ease of adding a single missing center to NPs with a large number of stereocenters. Indeed, analysing model performance across different NP classes showed that both full and partial assignment accuracies were particularly high for glycosides, nucleosides, lipids & terpenes which include steroids, and peptides, with **NPstereo** standing out again as the best performing model across most of these classes (**Table 7**).

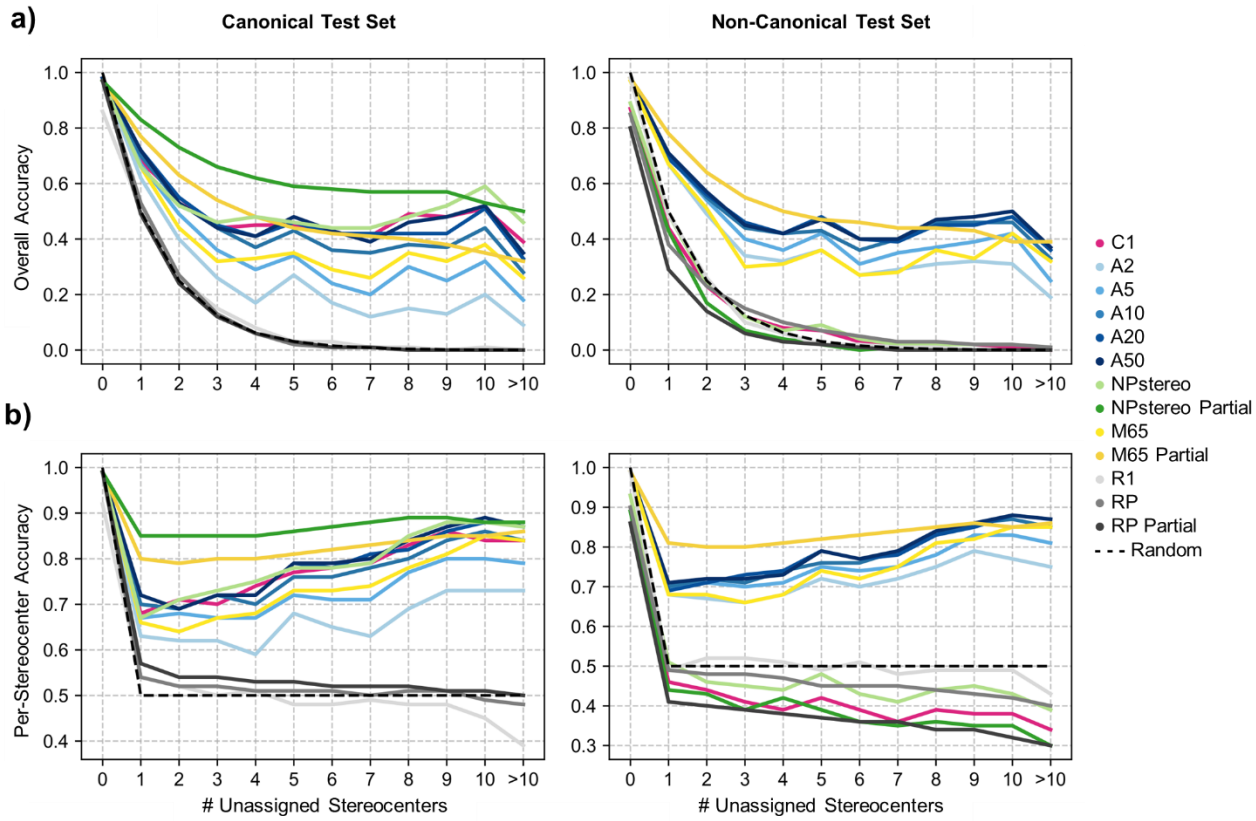


Figure 22. Model performance stratified by the number of unassigned stereocenters. (a) Top-1 NP stereochemistry full assignment accuracy for the canonical test set (left panel) and the non-canonical test set (right panel). (b) Top-1 NP stereochemistry per-stereocenter assignment accuracy for the canonical test set (left panel) and the non-canonical test set (right panel).

Table 7. Top-1 per-stereocenter accuracy for NP stereochemistry assignment, stratified by NP structural class. The best values for each NP structural class are highlighted in bold.

Model	Alkaloids (101)	Benzenoids (307)	Glycosides (872)	Lignans (63)	Lipids & Terpenes (1,674)	Nucleosides (15)	Peptides (298)	Polyketides (213)	Other (1,650)
	Top-1 accuracy (%) on full assignment canonical test set								
C1	82.4	69	84.8	70.2	79.5	92	80.7	71.4	73.4
A2	59.6	59.9	79.6	62.3	66.2	77	72.1	68	60.1
A5	65.2	60.5	83.2	69.4	73.7	89.2	75.2	76	67.8
A10	72.2	65.7	86.6	67.8	78.2	85.9	80.7	75.6	72
A20	77.6	**68.8**	88.4	65.2	80.3	**95.9**	83.3	76.1	73.4
A50	76.5	66.4	**88.9**	**73.6**	81.3	85.2	83.6	**78.1**	73.9
NPStereo	**83.6**	64.9	87.1	70.3	**81.9**	88.9	**83.7**	73.6	**74.3**
M65	67.8	59.5	86.8	66.6	75.3	93.3	79.8	71.5	68.8
R1	54.9	49.3	39.3	40.6	50.6	44	33.7	50.4	51.4
RP	52.4	48.7	49.2	50.5	49.9	46.1	49.5	53.2	52.2
	Top-1 accuracy (%) on partial assignment canonical test set								
NPStereo	**85.6**	**77.3**	**90.1**	**75.1**	**87.4**	**89.8**	**86**	**81.9**	**81.8**
M65	71.8	69.5	88.4	73.1	81.4	**89.8**	81.4	80.8	75.3
RP	53.4	55.3	52.8	62.2	53.5	49.3	52.4	55.4	53.2

7.2.4 Assigning stereochemistry with NPstereo

NPstereo might serve to assign the stereochemistry of partially or completely unassigned NPs. The assignment would be done by providing the query structure as a canonical SMILES, on which the model performed best. Here we illustrate how our model performed on such tasks with selected examples (**Figure 23**). First, we tested its performance for full assignment of stereochemistry on NP examples from the test set. For instance, **NPstereo** correctly assigned the stereochemistry of the well-known tubulin binding NPs colchicine (**1**),[303] docetaxel (**2**),[304] epothilone B (**3**),[305] and monomethyl auristatin E (**4**),[306] as well as the stereochemistry of both double bonds in bombykol (**5**), the first insect pheromone discovered.[307]

We further tested **NPstereo** with NPs with known stereochemistry, but which were only available as absolute SMILES in COCONUT and were therefore absent from model training. For example, **NPstereo** correctly assigned all stereocenters in the natural pyrethrin plant insecticide chrysanthemic acid (**6**),[308] the plant phenolic triterpenoid 1,3,6-tris-o-(3,4,5-trihydroxybenzoyl)hexopyranose (**7**),[309] the bacterial linear C_{30} carotenoid hydroxy-diaponeurosporenal (**8**),[310] and the plant NP D-α-tocopheryl acetate (vitamin E, **9**),[311] and only misassigned one stereocenter in the plant flavonoid glycoside datiscin (**10**).[312]

NPstereo also performed quite well when challenged to assign the stereochemistry of recently discovered NPs absent from the COCONUT dataset, such as the antibacterial fungal polyketide aspercitrininone A (**13**, all but one center correct),[270] antibacterial marine NP olimycin E (**11**, all centers correct),[313] the glucosylated plant alkaloid rhynchophylloside L 11-O-*β*-D-glucopyranoside (**12**, all centers correct),[314] the fungal polyketides aspercitrininone A (**13**, 4/5 stereocenters correct) [270] and mauritone A (**14**, all stereocenters correct).[315] Furthermore, **NPstereo** also correctly assigned all stereocenters in the plant neolignan (+)-nectamazin A (**15**), whose stereochemistry was recently reassigned.[316]

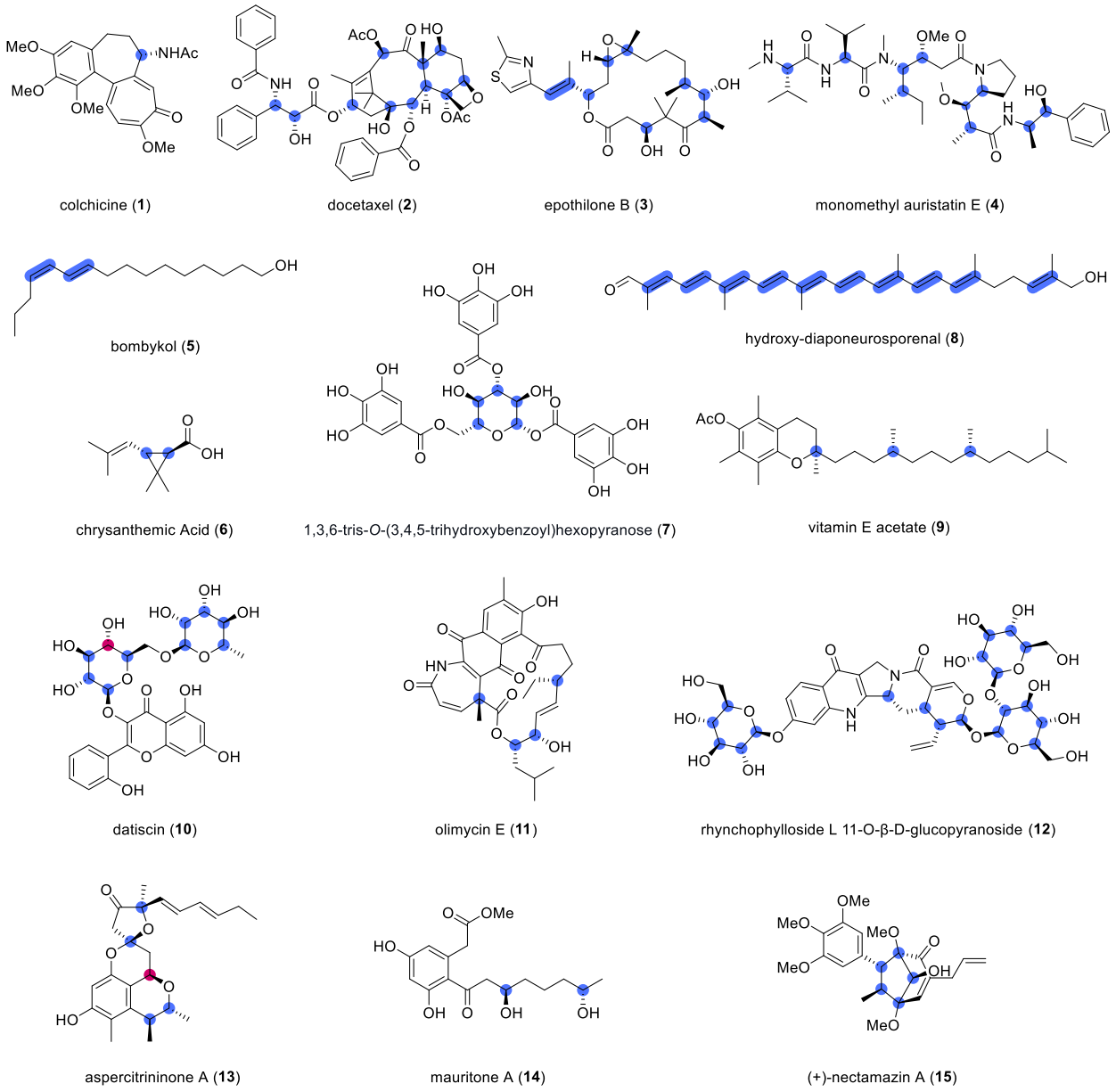


Figure 23. Assigning NP stereochemistry using NPstereo. The structural formulae of NPs with stereocenters highlighted as assigned correctly (blue) or incorrectly (red).

7.3 Conclusion

In this study, we demonstrated the efficacy of transformer-based models for assigning the stereochemistry of NPs from their absolute SMILES representations using data extracted from the COCONUT database. The selected model, named **NPstereo**, was trained and challenged with canonical SMILES, and achieved a per-stereocenter accuracy of 80.1% top-1 per-stereocenter accuracy for full stereochemical assignments and 86.3% top-1 per-stereocenter accuracy for partial

stereochemical assignments. The model showed a consistent ability to assign stereochemistry across molecules of varying complexity and performed particularly well for NPs with multiple stereocenters. Our work demonstrates that learnable stereochemical patterns exist in many NP classes and introduces a scalable methodology for assigning the stereochemistry of NPs, paving the way for future improvements in stereochemical prediction and NP characterization.

7.4 Methods

7.4.1 Dataset processing and visualization

The complete COCONUT database (09-2024) was retrieved from the website https://coconut.naturalproducts.net via the dedicated "Download" section as a PostreSQL dump. A SQL query was executed on the database to extract NPs with at least one associated citation DOI. The query retrieved the molecule identifier, canonical isomeric SMILES[171] structure and chemical class, for a total of 116,403 entries. The results were then exported as a CSV file for further processing.

The complete dataset of 116,403 NPs, consisting of isomeric SMILES with either fully or partially assigned stereocenters, were encoded using the MAP4C fingerprint.[42] The indices obtained from the MAP4C calculation were used to create a locality-sensitive hashing (LSH) forest of 32 trees. For each NP, the 20 approximate nearest neighbors in the MAP4C feature space were extracted from the LSH forest and used to calculate the TMAP layout. [39,41] The resulting layout was displayed in a static TMAP plot using the Python matplotlib package (3.5.3). The NPs in the TMAP were color-coded to highlight structural class, stereochemistry, and dataset split. Structural class information was extracted from the COCONUT "chemical_super_class" entry and further refined for peptides and glycosides using SMARTS patterns. The number of stereocenters was calculated using RDKit (2023.9.5).

7.4.2 Training data

NPs with incomplete stereochemistry (including both tetrahedral and double bond stereochemistry) were removed, reducing the dataset to 73,130 structures. These entries were then grouped by their canonical absolute SMILES (notation without stereochemistry), yielding 63,988 unique structures, each potentially associated with one or more stereoisomers. The canonical absolute SMILES dataset was divided into training, test, and validation sets using an 80:10:10 random split. In each of these three sets, each canonical absolute SMILES was associated with one or more, each written as a canonical isomeric SMILES. These canonical isomeric SMILES were used to form the target lists for training model **C1**.

To generate the corresponding source lists, we removed the stereochemical labels ("@" and "@@" for tetrahedral centers, "/C=C\" and "/C=C/" for double bonds) in each canonical isomeric SMILES to obtain an equivalent absolute SMILES. Importantly, different stereoisomers of the same canonical isomeric SMILES produced the same absolute SMILES after removing stereochemical labels, implying that the order of characters in the canonical isomeric SMILES did not contain stereochemical information. Also note that the absolute SMILES generated by removing stereochemical labels from canonical isomeric SMILES were almost identical to the canonical absolute SMILES, as measured by the Levenshtein distance between the two characters strings, compared to the Levenshtein distance between different absolute SMILES generated by SMILES randomization (**Figure E2**).

To obtain additional training data, we used several schemes of data augmentation in the training and validation sets separately. First, we applied SMILES randomization[300] to the canonical isomeric SMILES in the target lists of **C1** to increase their number by approximately factors of 2, 5, 10, 20, and 50 (after removal of duplicates), producing augmented target isomeric SMILES lists. We then generated the absolute SMILESs of each randomized isomeric SMILES by removing stereochemical labels for the source lists, resulting in training datasets for models **A2**, **A5**, **A10**, **A20** and **A50**. As a second data augmentation approach, we augmented the source lists of **C1** with SMILES containing only partially assigned stereochemistry. To do so, we identified the number of

stereocenters (n) in each molecule. For each molecule, we created up to 5 SMILES variations by randomly removing stereochemical labels for each level of stereochemistry removal, starting with removing all stereocenters (n), then n-1, and continuing until no stereocenters were replaced. This resulted in augmented source lists of isomeric SMILES strings with progressively reduced stereochemistry, each associated with the corresponding fully assigned canonical isomeric SMILES in the target lists, composing the training data for model **NPstereo**. In a third approach, we combined augmentation through SMILES randomization with augmentation by partial removal of stereochemical labels. To do so, we first augmented the target lists of **C1** 10-fold using SMILES randomization. In a second step, we augmented the source list by generating one additional variation for each level of stereochemistry removal for each randomized isomeric SMILES. This combined procedure resulted in approximately 65-fold data augmentation, providing training data for model **M65**.

Finally, we generated a first negative control training dataset **R1** by randomizing the stereochemical information in the target list of canonical isomeric SMILES of model **C1**, and a second negative control training dataset **RP** by augmenting the source lists of control **R1** using the partial assignment procedure used for model **NPstereo**.

For all datasets, isomeric SMILES (target) and the absolute SMILES generated from them (source) were tokenized using a custom tokenizer which applies a regular expression to split the SMILES string into individual chemical symbols, atoms, and bond types. The tokenizer captures elements like atoms (e.g., "Br", "Cl"), bond types (e.g., "=", "#"), and stereochemistry markers. All resulting training, validation, and test splits were saved as separate text files.

7.4.3 Transformer training

Model training was carried out on the OpenNMT python ecosystem (3.5.1).[317] All models used a transformer-based architecture with 6 layers each for both the encoder and decoder, employing 8 attention heads and a hidden size of 512. Training utilized mixed precision (fp16) and Adam

optimizer with a scheduled learning rate initialized at 2 and Noam decay. Batches were processed with a bucket size of 262,144 tokens and a batch size of 4096 tokens. Dropout regularization of 0.1 was applied during training, including attention dropout. The models were trained for a total of 100,000 steps. Checkpoints were saved every 25,000 steps, with validation performed every 5,000 steps to monitor model performance. The model hyperparameters and training parameters were configured according to the recommendations provided by OpenNMT. Complete configuration files for setup and training are available at https://github.com/reymond-group/NPstereo. Each model required approximately 9 hours to complete 100,000 training steps on a single Nvidia GeForce RTX 3070 GPU. The model checkpoint at step 100,000 was selected for subsequent performance evaluation across all trained models.

7.4.4 Performance evaluation

All calculations were done using the NumPy (1.26.4), pandas (2.1.0), and RDKit (2023.9.5) python libraries. The following performance metrics were used:

SMILES validity: Ratio of valid SMILES to the total number of predicted SMILES.

Full-Assignment Accuracy: Ratio of correctly predicted isomeric SMILES strings to the total number of predicted isomeric SMILES. An isomeric SMILES string predicted from an absolute SMILES is considered correctly predicted if it matches exactly one of the isomeric SMILES associated with this absolute SMILES.

Per-Stereocenter Accuracy: Average ratio of correctly predicted stereocenters within a single prediction, accounting for both tetrahedral stereocenters and stereogenic double bonds. When multiple associated stereoisomers are present, the highest ratio is used.

7.5 Code availability

The code for data extraction and augmentation, training the transformer models, running predictions, and analyzing results is available at https://github.com/reymond-group/NPstereo. The datasets used to train the models can be downloaded from https://zenodo.org/records/13790363.

7.6 Author Contribution Statement

MO designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper. Both authors read and approved the final manuscript.

7.7 Acknowledgements

This work was supported by the Swiss National Science Foundation (200020_178998) and the European Research Council (885076). MO thanks Yves Grandjean for his help with the transformer model implementation.

8 Conclusion and Outlook

8.1 Conclusion

While conventional molecular representations and predictive models are optimized for small organic molecules, they often fail to scale to larger and more complex structures. This thesis presented cheminformatics tools specifically designed to enable the comparison, generation, and classification of such structures.

The first part of the thesis investigated structural relationships between approved drugs using a reaction-informatics-based approach, where drug pairs are analysed through the lens of hypothetical transformations. Using DRFP and RXNMapper, molecular pairs were embedded into a reaction-like chemical space and ranked by atom-mapping confidence to distinguish feasible structural changes from non-trivial transmutations. This approach provided an orthogonal view of molecular similarity, enabling the assessment of scaffold modifications and substituent changes interesting for analog design.

To address the lack of stereochemistry-sensitive molecular fingerprints for larger compounds, this thesis introduced MAP4C. The fingerprint extends MAP4 by incorporating CIP stereochemistry annotations into circular substructures, allowing the differentiation of diastereomers and enantiomers across structurally diverse molecular sets. MAP4C performed on par with established fingerprints in virtual screening tasks and reliably distinguished thousands of stereoisomers in datasets of natural products and peptides, demonstrating its applicability for structure-based analysis of stereochemically rich molecules.

Building on these representation tools, this thesis introduced an updated implementation of the Peptide Design Genetic Algorithm (PDGA), capable of exploring peptide and peptoid spaces exceeding 10^60 structures. The algorithm was adapted to accommodate topologies compatible with solid-phase synthesis and feasible synthetic modifications. Its performance was demonstrated through target recovery benchmarks, where it successfully identified known antimicrobial peptides such as

polymyxin and cathelicidin, and through trajectory-based exploration yielding analogs with multiple sequence modifications but high structural similarity. The generated analogs populated coherent regions of chemical space and represent viable candidates for further study in structure-activity relationship investigations. The method was also applied in a real-life design scenario in collaboration with Dr. Etienne Bonvin, leading to the discovery of novel peptide-peptoid polymyxin analogs with experimentally confirmed antimicrobial activity.

This thesis also evaluated the use of general-purpose language models for molecular property prediction, focusing on GPT-3.5 as a case study. When applied to the classification of antimicrobial and hemolytic peptides, the model showed moderate baseline performance but was limited by poor reproducibility and high variability across runs. As a reference, a classical SVM trained on MAP4C consistently outperformed GPT-3.5 in both accuracy and reliability. These results suggest that while large language models may offer convenience and broad applicability, their current form is not well-suited for cheminformatics tasks requiring robustness, transparency, and controlled behaviour.

Finally, a transformer model (NPstereo) was developed for the stereochemical assignment of natural products from SMILES input lacking stereochemical labels. Trained on data extracted from the COCONUT database, the model achieved over 80% per-stereocenter accuracy in full assignment and over 86% in partial assignment tasks. These results indicated the presence of learnable patterns in NP stereochemistry and offer a machine learning-based alternative to wet-lab assignment experiments.

In summary, this thesis developed cheminformatics tools adapted to structurally complex molecules. The contributions addressed shortcomings in conventional methods and introduced practical solutions for comparison, generation, and classification tasks outside the scope of traditional small-molecule frameworks. The methods were designed to perform reliably in real-life scenarios, prioritizing consistency, modularity, and interpretability. All tools are available in open-source format and integrate with existing cheminformatics workflows, facilitating their application to evolving research and design problems involving peptides, natural products, and other large, diverse compound classes.

8.2 Outlook

The tools developed in this thesis were shaped by a practical need: to extend cheminformatics methods beyond the scope of small molecules and into the structurally rich space of peptides, natural products, and other complex scaffolds. As interest in these compound classes continues to grow, so does the importance of having efficient and scalable computational methods to support their design.

Although the molecular fingerprints presented here are broadly applicable, optimizing their computational performance could further improve the speed of similarity searches in large-scale virtual screening. Such improvements could be particularly relevant in early-stage screening campaigns, where rapid filtering across massive compound libraries is often a bottleneck. The Peptide Design Genetic Algorithm (PDGA) introduced in this work also offers a clear path for extension. While its current scoring relies on structural similarity to a reference molecule, future iterations could incorporate hybrid fitness functions that combine general scaffold constraints with machine learning-based property predictions. This would enable the algorithm to generate candidates guided not only by shape or topology, but also by modelled activity or selectivity, supporting more targeted and data-driven molecule generation.

On a broader level, the field continues to shift toward more flexible and modular compound classes such as peptides and peptidomimetics. These molecules challenge the assumptions built into conventional cheminformatics tools, but they also offer unique opportunities. Their synthetic accessibility, especially via solid-phase peptide synthesis, makes them particularly compatible with automation. This creates fertile ground for closed-loop design systems that combine *in silico* generation, predictive modelling, and experimental feedback. Current work within our group, in collaboration with Basak Olcay and Xiaoling Hu, is focused on building such automated frameworks for peptide discovery, aimed to streamline the design-make-test cycle.

In addition to these projects, I was also involved in a collaborative effort with Angelo Frei that applied machine learning to in-house experimental datasets (publication outlined in 1.2). The data originated from a focused screening of organometallic compounds for antibiotic activity and

represented a typical low-data regime. Despite this, it was possible to train predictive models that guided the selection of a second screening round, leading to a significantly higher hit rate in the second round. This experience demonstrated that even with limited data, machine learning can support decision-making in compound prioritization when paired with well-curated experimental inputs. It also reinforced the value of integrating predictive models into small-scale academic workflows, where iterative design is often guided by in-house knowledge and constraints.

Altogether, the contributions of this thesis offer a foundation for cheminformatics approaches tailored to emerging compound classes. As predictive models mature and synthesis becomes increasingly automated, there is clear potential for more integrated, iterative, and data-efficient design pipelines even in data constrained settings.

« So long, and thanks for all the fish. »

Douglas Adams, *So Long, and Thanks for All the Fish*

References

(1) Clackson, T.; Hoogenboom, H. R.; Griffiths, A. D.; Winter, G. Making Antibody Fragments Using Phage Display Libraries. *Nature* **1991**, *352* (6336), 624–628. https://doi.org/10.1038/352624a0.

(2) Winter, G.; Griffiths, A. D.; Hawkins, R. E.; Hoogenboom, H. R. Making Antibodies by Phage Display Technology. *Annu. Rev. Immunol.* **1994**, *12* (1), 433–455. https://doi.org/10.1146/annurev.iy.12.040194.002245.

(3) Brenner, S.; Lerner, R. A. Encoded Combinatorial Chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (12), 5381–5383. https://doi.org/10.1073/pnas.89.12.5381.

(4) Peterson, A. A.; Liu, D. R. Small-Molecule Discovery through DNA-Encoded Libraries. *Nat Rev Drug Discov* **2023**, *22* (9), 699–722. https://doi.org/10.1038/s41573-023-00713-6.

(5) Carle, V.; Kong, X.-D.; Comberlato, A.; Edwards, C.; Díaz-Perlas, C.; Heinis, C. Generation of a 100-Billion Cyclic Peptide Phage Display Library Having a High Skeletal Diversity. *Protein Engineering, Design and Selection* **2021**, *34*, gzab018. https://doi.org/10.1093/protein/gzab018.

(6) Huang, R. R.; Kierny, M.; Volgina, V.; Iwashima, M.; Miller, C.; Kay, B. K. Construction of an Ultra-Large Phage Display Library by Kunkel Mutagenesis and Rolling Circle Amplification. In *Phage Display*; Hust, M., Lim, T. S., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2023; Vol. 2702, pp 205–226. https://doi.org/10.1007/978-1-0716-3381-6_10.

(7) Cayley, E. Ueber Die Analytischen Figuren, Welche in Der Mathematik Bäume Genannt Werden Und Ihre Anwendung Auf Die Theorie Chemischer Verbindungen. *Ber. Dtsch. Chem. Ges.* **1875**, *8* (2), 1056–1059. https://doi.org/10.1002/cber.18750080252.

(8) Schiff, H. Zur Statistik Chemischer Verbindungen. *Ber. Dtsch. Chem. Ges.* **1875**, *8* (2), 1542–1547. https://doi.org/10.1002/cber.187500802191.

(9) Henze, H. R.; Blair, C. M. The Number Of Isoeric Hydrocarbons fF The Methane Series. *J. Am. Chem. Soc.* **1931**, *53* (8), 3077–3085. https://doi.org/10.1021/ja01359a034.

(10) Brinkmann, G.; Caporossi, G.; Hansen, P. A Survey and New Results on Computer Enumeration of Polyhex and Fusene Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 842–851. https://doi.org/10.1021/ci025526c.

(11) Dias, J. R. The Polyhex/Polypent Topological Paradigm: Regularities in the Isomer Numbers and Topological Properties of Select Subclasses of Benzenoid Hydrocarbons and Related Systems. *Chem. Soc. Rev.* **2010**, *39* (6), 1913. https://doi.org/10.1039/b913686j.

(12) 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.

(13) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432* (7019), 823–823. https://doi.org/10.1038/432823a.

(14) 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.

(15) Brown, F. K. Chemoinformatics: What Is It and How Does It Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*; Elsevier, 1998; Vol. 33, pp 375–384. https://doi.org/10.1016/S0065-7743(08)61100-8.

(16) Martinez-Mayorga, K.; Madariaga-Mazon, A.; Medina-Franco, J. L.; Maggiora, G. The Impact of Chemoinformatics on Drug Discovery in the Pharmaceutical Industry. *Expert Opinion on Drug Discovery* **2020**, *15* (3), 293–306. https://doi.org/10.1080/17460441.2020.1696307.

(17) Chen, H.; Kogej, T.; Engkvist, O. Cheminformatics in Drug Discovery, an Industrial Perspective. *Molecular Informatics* **2018**, *37* (9–10), 1800041. https://doi.org/10.1002/minf.201800041.
(18) Chen, Y.; Kirchmair, J. Cheminformatics in Natural Product-based Drug Discovery. *Molecular Informatics* **2020**, *39* (12), 2000171. https://doi.org/10.1002/minf.202000171.
(19) Zdrazil, B. Fifteen Years of ChEMBL and Its Role in Cheminformatics and Drug Discovery. *J Cheminform* **2025**, *17* (1), 32. https://doi.org/10.1186/s13321-025-00963-z.
(20) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat Rev Drug Discov* **2002**, *1* (11), 882–894. https://doi.org/10.1038/nrd941.
(21) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opinion on Drug Discovery* **2016**, *11* (2), 137–148. https://doi.org/10.1517/17460441.2016.1117070.
(22) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11* (23–24), 1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005.
(23) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996. https://doi.org/10.1021/ci9800211.
(24) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminf.* **2013**, *5* (1), 26. https://doi.org/10.1186/1758-2946-5-26.
(25) O'Boyle, N. M.; Sayle, R. A. Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. *J Cheminform* **2016**, *8* (1), 36. https://doi.org/10.1186/s13321-016-0148-0.
(26) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J Cheminform* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.
(27) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108–119. https://doi.org/10.1021/ci800249s.
(28) 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.
(29) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(30) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.
(31) Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and Applications of Chemical Fingerprint for Hit and Lead Screening. *Drug Discovery Today* **2022**, *27* (11), 103356. https://doi.org/10.1016/j.drudis.2022.103356.
(32) Probst, D.; Reymond, J.-L. A Probabilistic Molecular Fingerprint for Big Data Settings. *J Cheminform* **2018**, *10* (1), 66. https://doi.org/10.1186/s13321-018-0321-8.
(33) Broder, A. Z. On the Resemblance and Containment of Documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*; IEEE Comput. Soc: Salerno, Italy, 1998; pp 21–29. https://doi.org/10.1109/SEQUEN.1997.666900.
(34) 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*; ACM Press: Chiba, Japan, 2005; p 651. https://doi.org/10.1145/1060745.1060840.
(35) 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.
(36) 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.

(37) 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.
(38) Capecchi, A.; Awale, M.; Probst, D.; Reymond, J. PubChem and ChEMBL beyond Lipinski. *Mol. Inf.* **2019**, *38* (5), 1900016. https://doi.org/10.1002/minf.201900016.
(39) Orsi, M.; Probst, D.; Schwaller, P.; Reymond, J.-L. Alchemical Analysis of FDA Approved Drugs. *Digital Discovery* **2023**, *2* (5), 1289–1296. https://doi.org/10.1039/D3DD00039G.
(40) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J Cheminform* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.
(41) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J Cheminform* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.
(42) Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. *J Cheminform* **2024**, *16* (1), 53. https://doi.org/10.1186/s13321-024-00849-6.
(43) McConathy, J.; Owens, M. J. Stereochemistry in Drug Action. *Prim. Care Companion CNS Disord.* **2003**, *5* (2). https://doi.org/10.4088/PCC.v05n0202.
(44) Zheng, Y.; Mao, K.; Chen, S.; Zhu, H. Chirality Effects in Peptide Assembly Structures. *Front. Bioeng. Biotechnol.* **2021**, *9*, 703004. https://doi.org/10.3389/fbioe.2021.703004.
(45) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep Learning in Drug Discovery. *Molecular Informatics* **2016**, *35* (1), 3–14. https://doi.org/10.1002/minf.201501008.
(46) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.
(47) Hochreiter, S.; Klambauer, G.; Rarey, M. Machine Learning in Drug Discovery. *J. Chem. Inf. Model.* **2018**, *58* (9), 1723–1724. https://doi.org/10.1021/acs.jcim.8b00478.
(48) Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis*? *J. Chem. Inf. Model.* **2012**, *52* (6), 1413–1437. https://doi.org/10.1021/ci200409x.
(49) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach Learn* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.
(50) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
(51) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. In *Neurocomputing, Volume 1*; Anderson, J. A., Rosenfeld, E., Eds.; The MIT Press, 1988; pp 675–695. https://doi.org/10.7551/mitpress/4943.003.0128.
(52) Heikamp, K.; Bajorath, J. Support Vector Machines for Drug Discovery. *Expert Opinion on Drug Discovery* **2014**, *9* (1), 93–104. https://doi.org/10.1517/17460441.2014.866943.
(53) Rodríguez-Pérez, R.; Bajorath, J. Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery. *J Comput Aided Mol Des* **2022**, *36* (5), 355–362. https://doi.org/10.1007/s10822-022-00442-9.
(54) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J Comput Aided Mol Des* **2016**, *30* (8), 595–608. https://doi.org/10.1007/s10822-016-9938-8.
(55) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2020**, *63* (16), 8705–8722. https://doi.org/10.1021/acs.jmedchem.0c00385.
(56) Cai, H.; Zhang, H.; Zhao, D.; Wu, J.; Wang, L. FP-GNN: A Versatile Deep Learning Architecture for Enhanced Molecular Property Prediction. *Briefings in Bioinformatics* **2022**, *23* (6), bbac408. https://doi.org/10.1093/bib/bbac408.
(57) Atz, K.; Grisoni, F.; Schneider, G. Geometric Deep Learning on Molecular Representations. *Nat Mach Intell* **2021**, *3* (12), 1023–1032. https://doi.org/10.1038/s42256-021-00418-8.

(58) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17. https://doi.org/10.1021/acs.jcim.3c01250.
(59) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; Van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Commun Mater* **2022**, *3* (1), 93. https://doi.org/10.1038/s43246-022-00315-6.
(60) Yang, Z.; Chakraborty, M.; White, A. D. Predicting Chemical Shifts with Graph Neural Networks. *Chem. Sci.* **2021**, *12* (32), 10802–10809. https://doi.org/10.1039/D1SC01895G.
(61) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat Mach Intell* **2022**, *4* (3), 279–287. https://doi.org/10.1038/s42256-022-00447-x.
(62) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688-702.e13. https://doi.org/10.1016/j.cell.2020.01.021.
(63) Grisoni, F. Chemical Language Models for de Novo Drug Design: Challenges and Opportunities. *Current Opinion in Structural Biology* **2023**, *79*, 102527. https://doi.org/10.1016/j.sbi.2023.102527.
(64) Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; Engkvist, O. Reinvent 4: Modern AI–Driven Generative Molecule Design. *J Cheminform* **2024**, *16* (1), 20. https://doi.org/10.1186/s13321-024-00812-5.
(65) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1175–1183. https://doi.org/10.1021/acs.jcim.9b00943.
(66) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem. Sci.* **2021**, *12* (26), 9221–9232. https://doi.org/10.1039/D1SC01713F.
(67) 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.
(68) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv December 5, 2017. http://arxiv.org/abs/1706.03762 (accessed 2023-05-31).
(69) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv May 19, 2016. https://doi.org/10.48550/arXiv.1409.0473.
(70) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: A Pre-Trained Transformer for Computational Chemistry. *Mach. Learn.: Sci. Technol.* **2022**, *3* (1), 015022. https://doi.org/10.1088/2632-2153/ac3ffb.
(71) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736–8750. https://doi.org/10.1021/jacs.2c13467.
(72) Dollar, O.; Joshi, N.; Beck, D. A. C.; Pfaendtner, J. Attention-Based Generative Models for *de Novo* Molecular Design. *Chem. Sci.* **2021**, *12* (24), 8362–8372. https://doi.org/10.1039/D1SC01050F.
(73) Bran, A. M.; Schwaller, P. Transformers and Large Language Models for Chemistry and Drug Discovery. **2023**. https://doi.org/10.48550/ARXIV.2310.06083.
(74) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; De Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.;

Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digital Discovery* **2023**, *2* (5), 1233–1250. https://doi.org/10.1039/D3DD00113J.

(75) Chen, H.; Bajorath, J. Generative Design of Compounds with Desired Potency from Target Protein Sequences Using a Multimodal Biochemical Language Model. *J Cheminform* **2024**, *16* (1), 55. https://doi.org/10.1186/s13321-024-00852-x.

(76) Bajorath, J. Chemical Language Models for Molecular Design. *Molecular Informatics* **2024**, *43* (1), e202300288. https://doi.org/10.1002/minf.202300288.

(77) Duffy, B. C.; Zhu, L.; Decornez, H.; Kitchen, D. B. Early Phase Drug Discovery: Cheminformatics and Computational Techniques in Identifying Lead Series. *Bioorganic & Medicinal Chemistry* **2012**, *20* (18), 5324–5342. https://doi.org/10.1016/j.bmc.2012.04.062.

(78) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J Cheminform* **2021**, *13* (1), 12. https://doi.org/10.1186/s13321-020-00479-8.

(79) Stepišnik, T.; Škrlj, B.; Wicker, J.; Kocev, D. A Comprehensive Comparison of Molecular Feature Representations for Use in Predictive Modeling. *Computers in Biology and Medicine* **2021**, *130*, 104197. https://doi.org/10.1016/j.compbiomed.2020.104197.

(80) Van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

(81) Van Tilborg, D.; Rossen, L.; Grisoni, F. Molecular Deep Learning at the Edge of Chemical Space. March 17, 2025. https://doi.org/10.26434/chemrxiv-2025-qj4k3.

(82) Fooladi, H.; Vu, T. N. L.; Kirchmair, J. Evaluating Machine Learning Models for Molecular Property Prediction: Performance and Robustness on Out-of-Distribution Data. March 6, 2025. https://doi.org/10.26434/chemrxiv-2025-g1vjf-v2.

(83) Li, K.; Rubungo, A. N.; Lei, X.; Persaud, D.; Choudhary, K.; DeCost, B.; Dieng, A. B.; Hattrick-Simpers, J. Probing Out-of-Distribution Generalization in Machine Learning for Materials. *Commun Mater* **2025**, *6* (1), 9. https://doi.org/10.1038/s43246-024-00731-w.

(84) Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M. A. Transfer Learning: A Friendly Introduction. *J Big Data* **2022**, *9* (1), 102. https://doi.org/10.1186/s40537-022-00652-w.

(85) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63* (16), 8683–8694. https://doi.org/10.1021/acs.jmedchem.9b02147.

(86) King-Smith, E. Transfer Learning for a Foundational Chemistry Model. *Chem. Sci.* **2024**, *15* (14), 5143–5151. https://doi.org/10.1039/D3SC04928K.

(87) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes. *Org. Chem. Front.* **2021**, *8* (7), 1415–1423. https://doi.org/10.1039/D0QO01636E.

(88) Li, F.; Ackloo, S.; Arrowsmith, C. H.; Ban, F.; Barden, C. J.; Beck, H.; Beránek, J.; Berenger, F.; Bolotokova, A.; Bret, G.; Breznik, M.; Carosati, E.; Chau, I.; Chen, Y.; Cherkasov, A.; Corte, D. D.; Denzinger, K.; Dong, A.; Draga, S.; Dunn, I.; Edfeldt, K.; Edwards, A.; Eguida, M.; Eisenhuth, P.; Friedrich, L.; Fuerll, A.; Gardiner, S. S.; Gentile, F.; Ghiabi, P.; Gibson, E.; Glavatskikh, M.; Gorgulla, C.; Guenther, J.; Gunnarsson, A.; Gusev, F.; Gutkin, E.; Halabelian, L.; Harding, R. J.; Hillisch, A.; Hoffer, L.; Hogner, A.; Houliston, S.; Irwin, J. J.; Isayev, O.; Ivanova, A.; Jacquemard, C.; Jarrett, A. J.; Jensen, J. H.; Kireev, D.; Kleber, J.; Koby, S. B.; Koes, D.; Kumar, A.; Kurnikova, M. G.; Kutlushina, A.; Lessel, U.; Liessmann, F.; Liu, S.; Lu, W.; Meiler, J.; Mettu, A.; Minibaeva, G.; Moretti, R.; Morris, C. J.; Narangoda, C.; Noonan, T.; Obendorf, L.; Pach, S.; Pandit, A.; Perveen, S.; Poda, G.; Polishchuk, P.; Puls, K.; Pütter, V.; Rognan, D.; Roskams-Edris, D.; Schindler, C.; Sindt, F.; Spiwok, V.; Steinmann, C.; Stevens, R. L.; Talagayev, V.; Tingey, D.; Vu, O.; Walters, W. P.; Wang, X.; Wang, Z.; Wolber, G.; Wolf, C. A.; Wortmann, L.; Zeng, H.; Zepeda, C. A.; Zhang, K. Y. J.; Zhang, J.;

Zheng, S.; Schapira, M. CACHE Challenge #1: Targeting the WDR Domain of LRRK2, A Parkinson's Disease Associated Protein. *J. Chem. Inf. Model.* **2024**, *64* (22), 8521–8536. https://doi.org/10.1021/acs.jcim.4c01267.

(89) Moshawih, S.; Goh, H. P.; Kifli, N.; Idris, A. C.; Yassin, H.; Kotra, V.; Goh, K. W.; Liew, K. B.; Ming, L. C. Synergy between Machine Learning and Natural Products Cheminformatics: Application to the Lead Discovery of Anthraquinone Derivatives. *Chem Biol Drug Des* **2022**, *100* (2), 185–217. https://doi.org/10.1111/cbdd.14062.

(90) Tang, K. W. K.; Millar, B. C.; Moore, J. E. Antimicrobial Resistance (AMR). *Br J Biomed Sci* **2023**, *80*, 11387. https://doi.org/10.3389/bjbs.2023.11387.

(91) Dadgostar, P. Antimicrobial Resistance: Implications and Costs. *IDR* **2019**, *Volume 12*, 3903–3910. https://doi.org/10.2147/IDR.S234610.

(92) Bertagnolio, S.; Dobreva, Z.; Centner, C. M.; Olaru, I. D.; Donà, D.; Burzo, S.; Huttner, B. D.; Chaillon, A.; Gebreselassie, N.; Wi, T.; Hasso-Agopsowicz, M.; Allegranzi, B.; Sati, H.; Ivanovska, V.; Kothari, K. U.; Balkhy, H. H.; Cassini, A.; Hamers, R. L.; Weezenbeek, K. V.; Aanensen, D.; Alanio, A.; Alastruey-Izquierdo, A.; Alemayehu, T.; Al-Hasan, M.; Allegaert, K.; Al-Maani, A. S.; Al-Salman, J.; Alshukairi, A. N.; Amir, A.; Applegate, T.; Araj, G. F.; Villalobos, M. A.; Årdal, C.; Ashiru-Oredope, D.; Ashley, E. A.; Babin, F.-X.; Bachmann, L. H.; Bachmann, T.; Baker, K. S.; Balasegaram, M.; Bamford, C.; Baquero, F.; Barcelona, L. I.; Bassat, Q.; Bassetti, M.; Basu, S.; Beardsley, J.; Vásquez, G. B.; Berkley, J. A.; Bhatnagar, A. K.; Bielicki, J.; Bines, J.; Bongomin, F.; Bonomo, R. A.; Bradley, J. S.; Bradshaw, C.; Brett, A.; Brink, A.; Brown, C.; Brown, J.; Buising, K.; Carson, C.; Carvalho, A. C.; Castagnola, E.; Cavaleri, M.; Cecchini, M.; Chabala, C.; Chaisson, R. E.; Chakrabarti, A.; Chandler, C.; Chandy, S. J.; Charani, E.; Chen, L.; Chiara, F.; Chowdhary, A.; Chua, A.; Chuki, P.; Chun, D. R.; Churchyard, G.; Cirillo, D.; Clack, L.; Coffin, S. E.; Cohn, J.; Cole, M.; Conly, J.; Cooper, B.; Corso, A.; Cosgrove, S. E.; Cox, H.; Daley, C. L.; Darboe, S.; Darton, T.; Davies, G.; De Egea, V.; Dedeić-Ljubović, A.; Deeves, M.; Denkinger, C.; Dillon, J.-A. R.; Dramowski, A.; Eley, B.; Roberta Esposito, S. M.; Essack, S. Y.; Farida, H.; Farooqi, J.; Feasey, N.; Ferreyra, C.; Fifer, H.; Finlayson, H.; Frick, M.; Gales, A. C.; Galli, L.; Gandra, S.; Gerber, J. S.; Giske, C.; Gordon, B.; Govender, N.; Guessennd, N.; Guindo, I.; Gurbanova, E.; Gwee, A.; Hagen, F.; Harbarth, S.; Haze, J.; Heim, J.; Hendriksen, R.; Heyderman, R. S.; Holt, K. E.; Hönigl, M.; Hook, E. W.; Hope, W.; Hopkins, H.; Hughes, G.; Ismail, G.; Issack, M. I.; Jacobs, J.; Jasovský, D.; Jehan, F.; Pearson, A. J.; Jones, M.; Joshi, M. P.; Kapil, A.; Kariuki, S.; Karkey, A.; Kearns, G. L.; Keddy, K. H.; Khanna, N.; Kitamura, A.; Kolho, K.-L.; Kontoyiannis, D. P.; Kotwani, A.; Kozlov, R. S.; Kranzer, K.; Kularatne, R.; Lahra, M. M.; Langford, B. J.; Laniado-Laborin, R.; Larsson, D. G. J.; Lass-Flörl, C.; Le Doare, K.; Lee, H.; Lessa, F.; Levin, A. S.; Limmathurotsakul, D.; Lincopan, N.; Lo Vecchio, A.; Lodha, R.; Loeb, M.; Longtin, Y.; Lye, D. C.; Mahmud, A. M.; Manaia, C.; Manderson, L.; Mareković, I.; Marimuthu, K.; Martin, I.; Mashe, T.; Mei, Z.; Meis, J. F.; Lyra Tavares De Melo, F. A.; Mendelson, M.; Miranda, A. E.; Moore, D.; Morel, C.; Moremi, N.; Moro, M. L.; Moussy, F.; Mshana, S.; Mueller, A.; Ndow, F. J.; Nicol, M.; Nunn, A.; Obaro, S.; Obiero, C. W.; Okeke, I. N.; Okomo, U.; Okwor, T. J.; Oladele, R.; Omulo, S.; Ondoa, P.; Ortellado De Canese, J. M.; Ostrosky-Zeichner, L.; Padoveze, M. C.; Pai, M.; Park, B.; Parkhill, J.; Parry, C. M.; Peeling, R.; Sobreira Vieira Peixe, L. M.; Perovic, O.; Pettigrew, M. M.; Principi, N.; Pulcini, C.; Puspandari, N.; Rawson, T.; Reddy, D. L.; Reddy, K.; Redner, P.; Rodríguez Tudela, J. L.; Rodríguez-Baño, J.; Van Katwyk, S. R.; Roilides, E.; Rollier, C.; Rollock, L.; Ronat, J.-B.; Ruppe, E.; Sadarangani, M.; Salisbury, D.; Salou, M.; Samison, L. H.; Sanguinetti, M.; Sartelli, M.; Schellack, N.; Schouten, J.; Schwaber, M. J.; Seni, J.; Senok, A.; Shafer, W. M.; Shakoor, S.; Sheppard, D.; Shin, J.-H.; Sia, S.; Sievert, D.; Singh, I.; Singla, R.; Skov, R. L.; Soge, O. O.; Sprute, R.; Srinivasan, A.; Srinivasan, S.; Sundsfjord, A.; Tacconelli, E.; Tahseen, S.; Tangcharoensathien, V.; Tängdén, T.; Thursky, K.; Thwaites, G.; Tigulini De Souza Peral, R.; Tong, D.; Tootla, H. D.; Tsioutis, C.; Turner, K. M.; Turner, P.; Omar, S. V.; Van De Sande, W. W.; Van Den Hof, S.; Van Doorn, R.; Veeraraghavan, B.; Verweij, P.; Wahyuningsih, R.; Wang, H.; Warris, A.; Weinstock, H.; Wesangula, E.; Whiley, D.; White, P. J.; Williams, P.;

Xiao, Y.; Moscoso, M. Y.; Yang, H. L.; Yoshida, S.; Yu, Y.; Żabicka, D.; Zignol, M.; Rudan, I. WHO Global Research Priorities for Antimicrobial Resistance in Human Health. *The Lancet Microbe* **2024**, *5* (11), 100902. https://doi.org/10.1016/S2666-5247(24)00134-4.

(93) De Oliveira, D. M. P.; Forde, B. M.; Kidd, T. J.; Harris, P. N. A.; Schembri, M. A.; Beatson, S. A.; Paterson, D. L.; Walker, M. J. Antimicrobial Resistance in ESKAPE Pathogens. *Clin. Microbiol. Rev.* **2020**, *33* (3), 10.1128/cmr.00181-19. https://doi.org/10.1128/cmr.00181-19.

(94) Cook, M. A.; Wright, G. D. The Past, Present, and Future of Antibiotics. *Sci. Transl. Med.* **2022**, *14* (657), eabo7793. https://doi.org/10.1126/scitranslmed.abo7793.

(95) Roque-Borda, C. A.; Primo, L. M. D. G.; Franzyk, H.; Hansen, P. R.; Pavan, F. R. Recent Advances in the Development of Antimicrobial Peptides Against ESKAPE Pathogens. *Heliyon* **2024**, *0* (0). https://doi.org/10.1016/j.heliyon.2024.e31958.

(96) Scheenstra, M. R.; Belt, M. van den; Bokhoven, J. L. M. T.; Schneider, V. A. F.; Ordonez, S. R.; Dijk, A. van; Veldhuizen, E. J. A.; Haagsman, H. P. Cathelicidins PMAP-36, LL-37 and CATH-2 Are Similar Peptides with Different Modes of Action. *Sci. Rep.* **2019**, *9* (1), 1–12. https://doi.org/10.1038/s41598-019-41246-6.

(97) Poirel, L.; Jayol, A.; Nordmann, P. Polymyxins: Antibacterial Activity, Susceptibility Testing, and Resistance Mechanisms Encoded by Plasmids or Chromosomes. *Clin. Microbiol. Rev.* **2017**, *30* (2), 557–596. https://doi.org/10.1128/CMR.00064-16.

(98) 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.

(99) Zhang, Q.-Y.; Yan, Z.-B.; Meng, Y.-M.; Hong, X.-Y.; Shao, G.; Ma, J.-J.; Cheng, X.-R.; Liu, J.; Kang, J.; Fu, C.-Y. Antimicrobial Peptides: Mechanism of Action, Activity and Clinical Potential. *Military Med Res* **2021**, *8* (1), 48. https://doi.org/10.1186/s40779-021-00343-2.

(100) Orsi, M.; Reymond, J. Navigating a 1E+60 Chemical Space of Peptide/Peptoid Oligomers. *Molecular Informatics* **2024**, e202400186. https://doi.org/10.1002/minf.202400186.

(101) Geylan, G.; Janet, J. P.; Tibo, A.; He, J.; Patronov, A.; Kabeshov, M.; Czechtizky, W.; David, F.; Engkvist, O.; De Maria, L. PepINVENT: Generative Peptide Design beyond Natural Amino Acids. *Chem. Sci.* **2025**, 10.1039.D4SC07642G. https://doi.org/10.1039/D4SC07642G.

(102) 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.

(103) Bonvin, E.; Orsi, M.; Paschoud, T.; Gopalasingam, A.; Reusser, J.; Köhler, T.; Van Delden, C.; Reymond, J. Antimicrobial Peptide-Peptoid Macrocycles from the Polymyxin B2 Chemical Space. *Angewandte Chemie* **2025**, e202501299. https://doi.org/10.1002/ange.202501299.

(104) Orsi, M.; Reymond, J.-L. Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity? *RSC Med. Chem.* **2024**, *15* (6), 2030–2036. https://doi.org/10.1039/D4MD00159A.

(105) 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.

(106) 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.

(107) Zakharova, E.; Orsi, M.; Capecchi, A.; Reymond, J. Machine Learning Guided Discovery of Non-Hemolytic Membrane Disruptive Anticancer Peptides. *ChemMedChem* **2022**. https://doi.org/10.1002/cmdc.202200291.

(108) Wan, F.; De La Fuente-Nunez, C. Mining for Antimicrobial Peptides in Sequence Space. *Nat. Biomed. Eng* **2023**. https://doi.org/10.1038/s41551-023-01027-z.

(109) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3* (2), 157–166.

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

(111) Ivanenkov, Y. A.; Savchuk, N. P.; Ekins, S.; Balakin, K. V. Computational Mapping Tools for Drug Discovery. *Drug discovery today* **2009**, *14* (15–16), 767–775. http://dx.doi.org/10.1016/j.drudis.2009.05.016.
(112) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure−Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53* (23), 8209–8223. https://doi.org/10.1021/jm100933w.
(113) Sharma, S.; Arya, A.; Cruz, R.; Cleaves II, H. J. Automated Exploration of Prebiotic Chemical Reaction Space: Progress and Perspectives. *Life* **2021**, *11* (11), 1140. https://doi.org/10.3390/life11111140.
(114) Andronov, M.; Fedorov, M. V.; Sosnin, S. Exploring Chemical Reaction Space with Reaction Difference Fingerprints and Parametric T-SNE. *ACS Omega* **2021**, *6* (45), 30743–30751. https://doi.org/10.1021/acsomega.1c04778.
(115) Capecchi, A.; Reymond, J.-L. Classifying Natural Products from Plants, Fungi or Bacteria Using the COCONUT Database and Machine Learning. *J. Cheminf.* **2021**, *13* (1), 82. https://doi.org/10.1186/s13321-021-00559-3.
(116) Vriza, A.; Sovago, I.; Widdowson, D.; Kurlin, V.; A. Wood, P.; S. Dyer, M. Molecular Set Transformer: Attending to the Co-Crystals in the Cambridge Structural Database. *Digital Discovery* **2022**, *1* (6), 834–850. https://doi.org/10.1039/D2DD00068G.
(117) Medina-Franco, J. L.; Sánchez-Cruz, N.; López-López, E.; Díaz-Eufracio, B. I. Progress on Open Chemoinformatic Tools for Expanding and Exploring the Chemical Space. *J Comput Aided Mol Des* **2022**, *36* (5), 341–354. https://doi.org/10.1007/s10822-021-00399-1.
(118) Humer, C.; Heberle, H.; Montanari, F.; Wolf, T.; Huber, F.; Henderson, R.; Heinrich, J.; Streit, M. ChemInformatics Model Explorer (CIME): Exploratory Analysis of Chemical Model Explanations. *J Cheminform* **2022**, *14* (1), 21. https://doi.org/10.1186/s13321-022-00600-z.
(119) Beckers, M.; Fechner, N.; Stiefl, N. 25 Years of Small-Molecule Optimization at Novartis: A Retrospective Analysis of Chemical Series Evolution. *J. Chem. Inf. Model.* **2022**, *62* (23), 6002–6021. https://doi.org/10.1021/acs.jcim.2c00785.
(120) Zabolotna, Y.; Bonachera, F.; Horvath, D.; Lin, A.; Marcou, G.; Klimchuk, O.; Varnek, A. Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (18), 4537–4548. https://doi.org/10.1021/acs.jcim.2c00509.
(121) Cihan Sorkun, M.; Mullaj, D.; Koelman, J. M. V. A.; Er, S. ChemPlot, a Python Library for Chemical Space Visualization**. *Chemistry–Methods* **2022**, *2* (7), e202200005. https://doi.org/10.1002/cmtd.202200005.
(122) Han, M.; Liu, S.; Zhang, D.; Zhang, R.; Liu, D.; Xing, H.; Sun, D.; Gong, L.; Cai, P.; Tu, W.; Chen, J.; Hu, Q.-N. AddictedChem: A Data-Driven Integrated Platform for New Psychoactive Substance Identification. *Molecules* **2022**, *27* (12), 3931. https://doi.org/10.3390/molecules27123931.
(123) Moshawih, S.; Hadikhani, P.; Fatima, A.; Goh, H. P.; Kifli, N.; Kotra, V.; Goh, K. W.; Ming, L. C. Comparative Analysis of an Anthraquinone and Chalcone Derivatives-Based Virtual Combinatorial Library. A Cheminformatics "Proof-of-Concept" Study. *J. Mol. Graphics Modell.* **2022**, *117*, 108307. https://doi.org/10.1016/j.jmgm.2022.108307.
(124) 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.
(125) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (8), 3186–3204. https://doi.org/10.1021/jm401411z.
(126) J. Jesús Naveja; Medina-Franco, J. L. This File Contains the Six Compound Datasets Used in This Work in SDF Format, 2017. https://doi.org/10.5256/F1000RESEARCH.12095.D171632.
(127) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP. *Digital Discovery* **2022**, *1* (2), 91–97. https://doi.org/10.1039/D1DD00006C.

(128) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Science Advances* **2021**, *7* (15), eabe4166. https://doi.org/10.1126/sciadv.abe4166.
(129) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805.
(130) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. arXiv February 8, 2020. https://doi.org/10.48550/arXiv.1909.11942.
(131) Ball, P. Alchemical Culture and Poetry in Early Modern England. *Interdisciplinary Science Reviews* **2006**, *31* (1), 77–92. https://doi.org/10.1179/030801806X84246.
(132) Wentrup, C. Chemistry, Medicine, and Gold-Making: Tycho Brahe, Helwig Dieterich, Otto Tachenius, and Johann Glauber. *ChemPlusChem* **2023**, *88* (1), e202200289. https://doi.org/10.1002/cplu.202200289.
(133) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular Optimization by Capturing Chemist's Intuition Using Deep Neural Networks. *J. Cheminform.* **2021**, *13* (1), 26. https://doi.org/10.1186/s13321-021-00497-0.
(134) He, J.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Patronov, A.; Bjerrum, E. J.; Engkvist, O. Transformer-Based Molecular Optimization beyond Matched Molecular Pairs. *J. Cheminform.* **2022**, *14* (1), 18. https://doi.org/10.1186/s13321-022-00599-3.
(135) Lowe, Daniel. Chemical Reactions from US Patents (1976-Sep2016). *figshare. dataset.* **2017**. https://doi.org/10.6084/m9.figshare.5104873.v1.
(136) Kramer, C.; Fuchs, J. E.; Whitebread, S.; Gedeck, P.; Liedl, K. R. Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. *J. Med. Chem.* **2014**, *57* (9), 3786–3802. https://doi.org/10.1021/jm500317a.
(137) 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.
(138) 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.
(139) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technologies* **2004**, *1* (3), 217–224. https://doi.org/10.1016/j.ddtec.2004.10.009.
(140) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.
(141) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **2020**. https://doi.org/10.1021/acs.jcim.0c00675.
(142) 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.
(143) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.
(144) 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.
(145) Probst, D.; Reymond, J.-L. 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.

(146) Yang, Z.; Hackshaw, A.; Feng, Q.; Fu, X.; Zhang, Y.; Mao, C.; Tang, J. Comparison of Gefitinib, Erlotinib and Afatinib in Non-Small Cell Lung Cancer: A Meta-Analysis. *International Journal of Cancer* **2017**, *140* (12), 2805–2819. https://doi.org/10.1002/ijc.30691.
(147) Nordmann, P.; Poirel, L. Plasmid-Mediated Colistin Resistance: An Additional Antibiotic Resistance Menace. *Clin. Microbiol. Infect.* **2016**, *22* (5), 398–400. https://doi.org/10.1016/j.cmi.2016.03.009.
(148) Damashek, M. Gauging Similarity with N-Grams: Language-Independent Categorization of Text. *Science* **1995**, *267* (5199), 843–848. https://doi.org/10.1126/science.267.5199.843.
(149) 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.
(150) 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.
(151) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16* (8), 747–748. https://doi.org/10.1093/bioinformatics/16.8.747.
(152) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206. https://doi.org/10.1038/nbt1284.
(153) Czodrowski, P.; Bolick, W.-G. OCEAN: Optimized Cross rEActivity estimatioN. *J. Chem. Inf. Model.* **2016**, *56* (10), 2013–2023. https://doi.org/10.1021/acs.jcim.6b00067.
(154) 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.
(155) 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.
(156) Ertl, P.; Rohde, B. The Molecule Cloud - Compact Visualization of Large Collections of Molecules. *J. Cheminf.* **2012**, *4* (1), Article 12. http://www.jcheminf.com/content/4/1/12 (accessed Dec 6, 2012). https://doi.org/10.1186/1758-2946-4-12.
(157) 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.
(158) Ruddigkeit, L.; Blum, L. C.; Reymond, J. L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2013**, *53* (1), 56–65. https://doi.org/10.1021/ci300535x.
(159) 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.
(160) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of Chemical Space Networks Using a Tanimoto Similarity Variant Based upon Maximum Common Substructures. *J. Comput.-Aided Mol. Des.* **2015**, *29* (10), 937–950. https://doi.org/10.1007/s10822-015-9872-1.
(161) Blackmond, D. G. The Origin of Biological Homochirality. *Cold Spring Harb Perspect Biol* **2019**, *11* (3), a032540. https://doi.org/10.1101/cshperspect.a032540.
(162) Gal, J. Molecular Chirality in Chemistry and Biology: Historical Milestones. *Helv. Chim. Acta* **2013**, *96* (9), 1617–1657. https://doi.org/10.1002/hlca.201300300.
(163) Benner, S. A. Detecting Darwinism from Molecules in the Enceladus Plumes, Jupiter's Moons, and Other Planetary Water Lagoons. *Astrobiology* **2017**, *17* (9), 840–851. https://doi.org/10.1089/ast.2016.1611.
(164) H. Waldmann; Valeur, E.; Gueret, S. M.; Adihou, H.; Gopalakrishnan, R.; Lemurell, M.; Grossmann, T. N.; Plowright, A. T. New Modalities for Challenging Targets in Drug

Discovery. *Angew. Chem., Int. Ed. Engl.* **2017**, *56*, 10294–10323. https://doi.org/10.1002/anie.201611914.
(165) Caron, G.; Digiesi, V.; Solaro, S.; Ermondi, G. Flexibility in Early Drug Discovery: Focus on the beyond-Rule-of-5 Chemical Space. *Drug Discovery Today* **2020**. https://doi.org/10.1016/j.drudis.2020.01.012.
(166) Di Bonaventura, I.; Jin, X.; Visini, R.; Probst, D.; Javor, S.; Gan, B. H.; Michaud, G.; Natalello, A.; Doglia, S. M.; Kohler, T.; van 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.
(167) Cai, X.; Orsi, M.; Capecchi, A.; Köhler, T.; Delden, C. van; Javor, S.; Reymond, J.-L. An Intrinsically Disordered Antimicrobial Peptide Dendrimer from Stereorandomized Virtual Screening. *Cell Rep. Phys. Sci.* **2022**, *3* (12). https://doi.org/10.1016/j.xcrp.2022.101161.
(168) Personne, H.; Paschoud, T.; Fulgencio, S.; Baeriswyl, S.; Köhler, T.; van Delden, C.; Stocker, A.; Javor, S.; Reymond, J.-L. To Fold or Not to Fold: Diastereomeric Optimization of an α-Helical Antimicrobial Peptide. *J. Med. Chem.* **2023**, *66* (11), 7570–7583. https://doi.org/10.1021/acs.jmedchem.3c00460.
(169) 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.
(170) Manber, U. Finding Similar Files in a Large File System. In *Usenix Winter 1994 Technical Conference*; 1994; pp 1–10.
(171) 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.
(172) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.
(173) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminf.* **2021**, *13* (1), 2. https://doi.org/10.1186/s13321-020-00478-9.
(174) 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.
(175) 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.
(176) 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.
(177) 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.
(178) McGinnis, S.; Madden, T. L. BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic acids research* **2004**, *32* (Web Server issue), W20–W25. https://doi.org/10.1093/nar/gkh435.
(179) Dubikovskaya, E. A.; Thorne, S. H.; Pillow, T. H.; Contag, C. H.; Wender, P. A. Overcoming Multidrug Resistance of Small-Molecule Therapeutics through Conjugation with Releasable Octaarginine Transporters. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105* (34), 12128–12133. https://doi.org/10.1073/pnas.0805374105.
(180) Stanzl, E. G.; Trantow, B. M.; Vargas, J. R.; Wender, P. A. Fifteen Years of Cell-Penetrating, Guanidinium-Rich Molecular Transporters: Basic Science, Research Tools, and Clinical Applications. *Acc. Chem. Res.* **2013**, *46* (12), 2944–2954. https://doi.org/10.1021/ar4000554.

(181) 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.
(182) Buehler, Y.; Reymond, J.-L. Molecular Framework Analysis of the Generated Database GDB-13s. *J. Chem. Inf. Model.* **2023**, *63* (2), 484–492. https://doi.org/10.1021/acs.jcim.2c01107.
(183) Buehler, Y.; Reymond, J.-L. Expanding Bioactive Fragment Space with the Generated Database GDB-13s. *J. Chem. Inf. Model.* **2023**, *63* (20), 6239–6248. https://doi.org/10.1021/acs.jcim.3c01096.
(184) Lam, K. S.; Salmon, S. E.; Hersh, E. M.; Hruby, V. J.; Kazmierski, W. M.; Knapp, R. J. A New Type of Synthetic Peptide Library for Identifying Ligand-Binding Activity. *Nature* **1991**, *354* (6348), 82–84. https://doi.org/10.1038/354082a0.
(185) Houghten, R. A.; Pinilla, C.; Blondelle, S. E.; Appel, J. R.; Dooley, C. T.; Cuervo, J. H. Generation and Use of Synthetic Peptide Combinatorial Libraries for Basic Research and Drug Discovery. *Nature* **1991**, *354* (6348), 84–86. https://doi.org/10.1038/354084a0.
(186) Lam, K. S.; Lebl, M.; Krchňák, V. The "One-Bead-One-Compound" Combinatorial Library Method. *Chem. Rev.* **1997**, *97* (2), 411–448. https://doi.org/10.1021/cr9600114.
(187) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2* (5), 369–378.
(188) Glökler, J.; Schütze, T.; Konthur, Z. Automation in the High-Throughput Selection of Random Combinatorial Libraries—Different Approaches for Select Applications. *Molecules* **2010**, *15* (4), 2478–2490. https://doi.org/10.3390/molecules15042478.
(189) Goto, Y.; Suga, H. The RaPID Platform for the Discovery of Pseudo-Natural Macrocyclic Peptides. *Acc. Chem. Res.* **2021**, *54* (18), 3604–3617. https://doi.org/10.1021/acs.accounts.1c00391.
(190) Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol. Transl. Sci.* **2021**, *4* (4), 1265–1279. https://doi.org/10.1021/acsptsci.1c00118.
(191) Dockerill, M.; Winssinger, N. DNA-Encoded Libraries: Towards Harnessing Their Full Power with Darwinian Evolution. *Angewandte Chemie* **2023**, *135* (9), e202215542. https://doi.org/10.1002/ange.202215542.
(192) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432* (7019), 855–861. https://doi.org/10.1038/nature03193.
(193) Renner, S.; van Otterlo, W. A. L.; Dominguez Seoane, M.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-Guided Mapping and Navigation of Chemical Space. *Nat. Chem. Biol.* **2009**, *5* (8), 585–592. https://doi.org/10.1038/nchembio.188.
(194) Bon, M.; Bilsland, A.; Bower, J.; McAulay, K. Fragment-Based Drug Discovery—the Importance of High-Quality Molecule Libraries. *Mol. Oncol.* **2022**, *16* (21), 3761–3777. https://doi.org/10.1002/1878-0261.13277.
(195) Di Bonaventura, I.; Baeriswyl, S.; Capecchi, A.; Gan, B.-H.; Jin, X.; Siriwardena, T. N.; He, R.; Kohler, T.; Pompilio, A.; Di Bonaventura, G.; van Delden, C.; Javor, S.; Reymond, J.-L. An Antimicrobial Bicyclic Peptide from Chemical Space Against Multidrug Resistant Gram-Negative Bacteria. *Chem. Commun.* **2018**, *54*, 5130–5133. https://doi.org/10.1039/c8cc02412j.
(196) Merz, M. L.; Habeshian, S.; Li, B.; David, J.-A. G.; Nielsen, A. L.; Ji, X.; Il Khwildy, K.; Duany Benitez, M. M.; Phothirath, P.; Heinis, C. De Novo Development of Small Cyclic Peptides That Are Orally Bioavailable. *Nat. Chem. Biol.* **2023**, 1–10.
(197) Chevillard, F.; Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *Journal of chemical information and modeling* **2015**, *55* (9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.

(198) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.
(199) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681. https://doi.org/10.1016/j.isci.2020.101681.
(200) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021–2034. https://doi.org/10.1021/acs.jcim.2c00224.
(201) Irwin, J. J.; Gaskins, G.; Sterling, T.; Mysinger, M. M.; Keiser, M. J. Predicted Biological Activity of Purchasable Chemical Space. *Journal of chemical information and modeling* **2018**, *58* (1), 148–164. https://doi.org/10.1021/acs.jcim.7b00316.
(202) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229. https://doi.org/10.1038/s41586-019-0917-9.
(203) Korn, M.; Ehrt, C.; Ruggiu, F.; Gastreich, M.; Rarey, M. Navigating Large Chemical Spaces in Early-Phase Drug Discovery. *Curr. Opin. Struct. Biol.* **2023**, *80*, 102578. https://doi.org/10.1016/j.sbi.2023.102578.
(204) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Prot. Sci.* **1998**, *7* (9), 1884–1897. https://doi.org/10.1002/pro.5560070905.
(205) Kandel, J.; Tayara, H.; Chong, K. T. PUResNet: Prediction of Protein-Ligand Binding Sites Using Deep Residual Neural Network. *J. Cheminform.* **2021**, *13* (1), 65. https://doi.org/10.1186/s13321-021-00547-7.
(206) Comajuncosa-Creus, A.; Jorba, G.; Barril, X.; Aloy, P. Comprehensive Detection and Characterization of Human Druggable Pockets through Novel Binding Site Descriptors. bioRxiv March 16, 2024, p 2024.03.14.584971. https://doi.org/10.1101/2024.03.14.584971.
(207) Reymond, J.-L.; Ruddigkeit, L.; Blum, L.; Deursen, R. van. The Enumeration of Chemical Space. *WIREs Comput. Mol. Sci.* **2012**, *2* (5), 717–733. https://doi.org/10.1002/wcms.1104.
(208) Awale, M.; Visini, R.; Probst, D.; Arus-Pous, J.; Reymond, J. L. Chemical Space: Big Data Challenge for Molecular Diversity. *Chimia* **2017**, *71* (10), 661–666. https://doi.org/10.2533/chimia.2017.661.
(209) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discovery Today* **2011**, *16* (9), 372–376. https://doi.org/10.1016/j.drudis.2011.02.011.
(210) Giordano, D.; Biancaniello, C.; Argenio, M. A.; Facchiano, A. Drug Design by Pharmacophore and Virtual Screening Approach. *Pharmaceuticals* **2022**, *15* (5), 646. https://doi.org/10.3390/ph15050646.
(211) Schmidt, R.; Klein, R.; Rarey, M. Maximum Common Substructure Searching in Combinatorial Make-on-Demand Compound Spaces. *J. Chem. Inf. Model.* **2022**, *62* (9), 2133–2150. https://doi.org/10.1021/acs.jcim.1c00640.
(212) 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.
(213) Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminf.* **2015**, *7*, 3. noc
(214) Bonvin, E.; Personne, H.; Paschoud, T.; Reusser, J.; Gan, B.-H.; Luscher, A.; Köhler, T.; van Delden, C.; Reymond, J.-L. Antimicrobial Peptide–Peptoid Hybrids with and without Membrane Disruption. *ACS Infect. Dis.* **2023**, *9* (12), 2593–2606. https://doi.org/10.1021/acsinfecdis.3c00421.
(215) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663.

(216) Lamanna, G.; Delre, P.; Marcou, G.; Saviano, M.; Varnek, A.; Horvath, D.; Mangiatordi, G. F. GENERA: A Combined Genetic/Deep-Learning Algorithm for Multiobjective Target-Oriented De Novo Design. *J. Chem. Inf. Model.* **2023**, *63* (16), 5107–5119. https://doi.org/10.1021/acs.jcim.3c00963.
(217) Cai, X.; Capecchi, A.; Olcay, B.; Orsi, M.; Javor, S.; Reymond, J.-L. Exploring the Sequence Space of Antimicrobial Peptide Dendrimers. *Isr. J. Chem.* **2023**, *63* (10–11), e202300096. https://doi.org/10.1002/ijch.202300096.
(218) Zuckermann, R. N. Peptoid Origins. *Peptide Sci.* **2011**, *96* (5), 545–555. https://doi.org/10.1002/bip.21573.
(219) Amblard, M.; Fehrentz, J.-A.; Martinez, J.; Subra, G. Methods and Protocols of Modern Solid Phase Peptide Synthesis. *Mol. Biotechnol.* **2006**, *33* (3), 239–254. https://doi.org/10.1385/MB:33:3:239.
(220) Clapperton, A. M.; Babi, J.; Tran, H. A Field Guide to Optimizing Peptoid Synthesis. *ACS Polym. Au* **2022**, *2* (6), 417–429. https://doi.org/10.1021/acspolymersau.2c00036.
(221) Matthews, T.; Salgo, M.; Greenberg, M.; Chung, J.; DeMasi, R.; Bolognesi, D. Enfuvirtide: The First Therapy to Inhibit the Entry of HIV-1 into Host CD4 Lymphocytes. *Nat. Rev. Drug Discov.* **2004**, *3* (3), 215–225. https://doi.org/10.1038/nrd1331.
(222) Knudsen, L. B.; Lau, J. The Discovery and Development of Liraglutide and Semaglutide. *Front. Endocrinol.* **2019**, *10*. https://doi.org/10.3389/fendo.2019.00155.
(223) Zuckermann, R. N.; Kerr, J. M.; Kent, S. B. H.; Moos, W. H. Efficient Method for the Preparation of Peptoids [Oligo(N-Substituted Glycines)] by Submonomer Solid-Phase Synthesis. *J. Am. Chem. Soc.* **1992**, *114* (26), 10646–10647. https://doi.org/10.1021/ja00052a076.
(224) Morstein, J.; Capecchi, A.; Hinnah, K.; Park, B.; Petit-Jacques, J.; Van Lehn, R. C.; Reymond, J.-L.; Trauner, D. Medium-Chain Lipid Conjugation Facilitates Cell-Permeability and Bioactivity. *J. Am. Chem. Soc.* **2022**, *144* (40), 18532–18544. https://doi.org/10.1021/jacs.2c07833.
(225) Kurtzhals, P.; Havelund, S.; Jonassen, I.; Markussen, J. Effect of Fatty Acids and Selected Drugs on the Albumin Binding of a Long-Acting, Acylated Insulin Analogue. *Journal of pharmaceutical sciences* **1997**, *86* (12), 1365–1368. https://doi.org/10.1021/js9701768.
(226) 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 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956.
(227) Gause, G. F.; Brazhnikova, M. G. Gramicidin S and Its Use in the Treatment of Infected Wounds. *Nature* **1944**, *154* (3918), 703–703. https://doi.org/10.1038/154703a0.
(228) Kondejewski, L. H.; Farmer, S. W.; Wishart, D. S.; Hancock, R. E. w.; Hodges, R. S. Gramicidin S Is Active against Both Gram-Positive and Gram-Negative Bacteria. *International Journal of Peptide and Protein Research* **1996**, *47* (6), 460–466. https://doi.org/10.1111/j.1399-3011.1996.tb01096.x.
(229) Knappe, D.; Piantavigna, S.; Hansen, A.; Mechler, A.; Binas, A.; Nolte, O.; Martin, L. L.; Hoffmann, R. Oncocin (VDKPPYLPRPRPPRRIYNR-NH2): A Novel Antibacterial Peptide Optimized against Gram-Negative Human Pathogens. *J. Med. Chem.* **2010**, *53* (14), 5240–5247. https://doi.org/10.1021/jm100378b.
(230) Zhang, H.; Xia, X.; Han, F.; Jiang, Q.; Rong, Y.; Song, D.; Wang, Y. Cathelicidin-BF, a Novel Antimicrobial Peptide from *Bungarus Fasciatus* , Attenuates Disease in a Dextran Sulfate Sodium Model of Colitis. *Mol. Pharmaceutics* **2015**, *12* (5), 1648–1661. https://doi.org/10.1021/acs.molpharmaceut.5b00069.
(231) Bokesch, H. R.; Pannell, L. K.; Cochran, P. K.; Sowder, R. C.; McKee, T. C.; Boyd, M. R. A Novel Anti-HIV Macrocyclic Peptide from Palicourea Condensata. *J. Nat. Prod.* **2001**, *64* (2), 249–250. https://doi.org/10.1021/np000372l.
(232) Brown, N.; McKay, B.; Gasteiger, J. The de Novo Design of Median Molecules within a Property Range of Interest. *J. Comput.-Aided Mol. Des.* **2004**, *18* (12), 761–771.

(233) van Deursen, R.; Reymond, J.-L. Chemical Space Travel. *ChemMedChem* **2007**, *2* (5), 636–640. https://doi.org/10.1002/cmdc.200700021.
(234) Chatterjee, J.; Rechenmacher, F.; Kessler, H. N-Methylation of Peptides and Proteins: An Important Element for Modulating Biological Functions. *Angew. Chem., Int. Ed. Engl.* **2013**, *52* (1), 254–269. https://doi.org/10.1002/anie.201205674.
(235) Corbett, K. M.; Ford, L.; Warren, D. B.; Pouton, C. W.; Chalmers, D. K. Cyclosporin Structure and Permeability: From A to Z and Beyond. *J. Med. Chem.* **2021**, *64* (18), 13131–13151. https://doi.org/10.1021/acs.jmedchem.1c00580.
(236) Lakemeyer, M.; Zhao, W.; Mandl, F. A.; Hammann, P.; Sieber, S. A. Thinking Outside the Box-Novel Antibacterials To Tackle the Resistance Crisis. *Angew. Chem. Int. Ed.* **2018**, *57* (44), 14440–14475. https://doi.org/10.1002/anie.201804971.
(237) 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. *The Lancet Infectious Diseases* **2020**, *20* (9), e216–e230. https://doi.org/10.1016/S1473-3099(20)30327-3.
(238) Mookherjee, N.; Anderson, M. A.; Haagsman, H. P.; Davidson, D. J. Antimicrobial Host Defence Peptides: Functions and Clinical Potential. *Nat Rev Drug Discov* **2020**, *19* (5), 311–332. https://doi.org/10.1038/s41573-019-0058-8.
(239) Torres, M. D. T.; Sothiselvam, S.; Lu, T. K.; De La Fuente-Nunez, C. Peptide Design Principles for Antimicrobial Applications. *Journal of Molecular Biology* **2019**, *431* (18), 3547–3567. https://doi.org/10.1016/j.jmb.2018.12.015.
(240) Capecchi, A.; Reymond, J.-L. Peptides in Chemical Space. *Med. Drug Discovery* **2021**, *9*, 100081. https://doi.org/10.1016/j.medidd.2021.100081.
(241) Liu, S. Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides. *Scientific Reports* **2018**.
(242) 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.
(243) 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. **2019**.
(244) 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. *Molecular Therapy - Nucleic Acids* **2020**, *20*, 882–894. https://doi.org/10.1016/j.omtn.2020.05.006.
(245) Liu, G.; Catacutan, D. B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J. C.; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, J. M. Deep Learning-Guided Discovery of an Antibiotic Targeting Acinetobacter Baumannii. *Nat Chem Biol* **2023**. https://doi.org/10.1038/s41589-023-01349-8.
(246) Aguilera-Puga, M. D. C.; Plisson, F. *Structure-Aware Machine Learning Strategies for Antimicrobial Peptide Discovery*; preprint; In Review, 2024. https://doi.org/10.21203/rs.3.rs-3938402/v1.
(247) Wan, F.; Wong, F.; Collins, J. J.; De La Fuente-Nunez, C. Machine Learning for Antimicrobial Peptide Identification and Design. *Nat Rev Bioeng* **2024**. https://doi.org/10.1038/s44222-024-00152-x.
(248) 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.
(249) Hasan, M. M.; Schaduangrat, N.; Basith, S.; Lee, G.; Shoombuatong, W.; Manavalan, B. HLPpred-Fuse: Improved and Robust Prediction of Hemolytic Peptide and Its Activity by

Fusing Multiple Feature Representation. *Bioinformatics* **2020**, *36* (11), 3350–3356. https://doi.org/10.1093/bioinformatics/btaa160.
(250) Ansari, M.; White, A. D. Serverless Prediction of Peptide Properties with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2023**.
(251) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
(252) Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv October 7, 2014. http://arxiv.org/abs/1409.1259 (accessed 2023-05-31).
(253) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. arXiv July 22, 2020. http://arxiv.org/abs/2005.14165 (accessed 2023-05-31).
(254) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging Large Language Models for Predictive Chemistry. *Nat Mach Intell* **2024**, *6* (2), 161–169. https://doi.org/10.1038/s42256-023-00788-1.
(255) Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What Can Large Language Models Do in Chemistry? A Comprehensive Benchmark on Eight Tasks.
(256) Castro Nascimento, C. M.; Pimentel, A. S. Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *J. Chem. Inf. Model.* **2023**, *63* (6), 1649–1655. https://doi.org/10.1021/acs.jcim.3c00285.
(257) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Peña Ccoa, W. J. Assessment of Chemistry Knowledge in Large Language Models That Generate Code. *Digital Discovery* **2023**, *2* (2), 368–376. https://doi.org/10.1039/D2DD00087C.
(258) Bran, A. M.; Cox, S.; White, A. D.; Schwaller, P. ChemCrow: Augmenting Large-Language Models with Chemistry Tools. arXiv April 12, 2023. http://arxiv.org/abs/2304.05376 (accessed 2023-05-31).
(259) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* **2023**, *624* (7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0.
(260) 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.
(261) 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.
(262) 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.
(263) 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.
(264) Mori, K. Bioactive Natural Products and Chirality. *Chirality* **2011**, *23* (6), 449–462. https://doi.org/10.1002/chir.20930.
(265) Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat Rev Drug Discov* **2021**, *20* (3), 200–216. https://doi.org/10.1038/s41573-020-00114-z.
(266) Huang, M.; Lu, J.-J.; Ding, J. Natural Products in Cancer Therapy: Past, Present and Future. *Nat. Prod. Bioprospect.* **2021**, *11* (1), 5–13. https://doi.org/10.1007/s13659-020-00293-7.

(267) Porras, G.; Chassagne, F.; Lyles, J. T.; Marquez, L.; Dettweiler, M.; Salam, A. M.; Samarakoon, T.; Shabih, S.; Farrokhi, D. R.; Quave, C. L. Ethnobotany and the Role of Plant Natural Products in Antibiotic Drug Discovery. *Chem. Rev.* **2021**, *121* (6), 3495–3560. https://doi.org/10.1021/acs.chemrev.0c00922.
(268) Dai, J.-M.; Yan, B.-C.; Hu, K.; Li, X.-R.; Li, X.; Sun, H.-D.; Puno, P.-T. Isoxerophilusins A and B, Two Novel Polycyclic Asymmetric Diterpene Dimers from *Isodon Xerophilus* : Structural Elucidation, Modification, and Inhibitory Activities against α-Glucosidase. *Org. Lett.* **2024**, *26* (29), 6203–6208. https://doi.org/10.1021/acs.orglett.4c02095.
(269) Wang, Q.-Y.; Gao, Y.; Yao, J.-N.; Zhou, L.; Chen, H.-P.; Liu, J.-K. Penisimplicins A and B: Novel Polyketide–Peptide Hybrid Alkaloids from the Fungus Penicillium Simplicissimum JXCC5. *Molecules* **2024**, *29* (3), 613. https://doi.org/10.3390/molecules29030613.
(270) Li, T.-X.; Dong, H.-H.; Xing, L.; He, L.; Zhang, R.-Y.; Shao, D.-Y.; Dai, Y.-X.; Li, D.-L.; Xu, C.-P. Aspercitrininone A, Novel Antibacterial Polyketide Featuring Unusual Spiral Skeleton from Aspergillus Cristatus. *Fitoterapia* **2024**, *173*, 105827. https://doi.org/10.1016/j.fitote.2024.105827.
(271) Sun, J.; Ma, J.; Zhang, S.; Zhang, L.; Tao, X.; Li, C.; Zang, Y.; Ji, M.; Tao, A.; Zhang, D. Magterpenes A–C: Three Meroterpenoids with an Unprecedented 6/6/6/6/6 Polycyclic Skeleton Extracted from *Magnolia Officinalis* Rehd. et Wils. *J. Org. Chem.* **2024**, *89* (12), 8871–8877. https://doi.org/10.1021/acs.joc.4c00739.
(272) Xu, Y.; Zhang, Y.; Zhang, Q.; Li, J. C.; Zhou, Z. H.; Yang, Z.; Xiu, J.; Chen, X.; Huang, J.; Ge, H. M.; Shi, J. Genome Mining of Cinnamoyl-Containing Nonribosomal Peptide Gene Clusters Directs the Production of Malacinnamycin. *Org. Lett.* **2024**, acs.orglett.4c00052. https://doi.org/10.1021/acs.orglett.4c00052.
(273) Hagar, M.; Morgan, K. D.; Stumpf, S. D.; Tsingos, M.; Banuelos, C. A.; Sadar, M. D.; Blodgett, J. A. V.; Andersen, R. J.; Ryan, K. S. Piperazate-Guided Isolation of Caveamides A and B, Cyclohexenylalanine-Containing Nonribosomal Peptides from a Cave Actinomycete. *Org. Lett.* **2024**, *26* (19), 4127–4131. https://doi.org/10.1021/acs.orglett.4c01218.
(274) Arishi, A. A.; Shang, Z.; Lacey, E.; Crombie, A.; Vuong, D.; Li, H.; Bracegirdle, J.; Turner, P.; Lewis, W.; Flematti, G. R.; Piggott, A. M.; Chooi, Y.-H. Discovery and Heterologous Biosynthesis of Glycosylated Polyketide Luteodienoside A Reveals Unprecedented Glucinol-Mediated Product Offloading by a Fungal Carnitine *O* -Acyltransferase Domain. *Chem. Sci.* **2024**, *15* (9), 3349–3356. https://doi.org/10.1039/D3SC05008D.
(275) Young, R. J.; Flitsch, S. L.; Grigalunas, M.; Leeson, P. D.; Quinn, R. J.; Turner, N. J.; Waldmann, H. The Time and Place for Nature in Drug Discovery. *JACS Au* **2022**, *2* (11), 2400–2416. https://doi.org/10.1021/jacsau.2c00415.
(276) Gil, R. R. Constitutional, Configurational, and Conformational Analysis of Small Organic Molecules on the Basis of NMR Residual Dipolar Couplings. *Angew Chem Int Ed* **2011**, *50* (32), 7222–7224. https://doi.org/10.1002/anie.201101561.
(277) Liu, Y.; Saurí, J.; Mevers, E.; Peczuh, M. W.; Hiemstra, H.; Clardy, J.; Martin, G. E.; Williamson, R. T. Unequivocal Determination of Complex Molecular Structures Using Anisotropic NMR Measurements. *Science* **2017**, *356* (6333), eaam5349. https://doi.org/10.1126/science.aam5349.
(278) Liu, Y.; Navarro-Vázquez, A.; Gil, R. R.; Griesinger, C.; Martin, G. E.; Williamson, R. T. Application of Anisotropic NMR Parameters to the Confirmation of Molecular Structure. *Nat Protoc* **2019**, *14* (1), 217–247. https://doi.org/10.1038/s41596-018-0091-9.
(279) Marcarino, M. O.; Zanardi, M. M.; Cicetti, S.; Sarotti, A. M. NMR Calculations with Quantum Methods: Development of New Tools for Structural Elucidation and Beyond. *Acc. Chem. Res.* **2020**, *53* (9), 1922–1932. https://doi.org/10.1021/acs.accounts.0c00365.
(280) Polavarapu, P. L.; Santoro, E. Vibrational Optical Activity for Structural Characterization of Natural Products. *Nat. Prod. Rep.* **2020**, *37* (12), 1661–1699. https://doi.org/10.1039/D0NP00025F.
(281) Pescitelli, G. ECD Exciton Chirality Method Today: A Modern Tool for Determining Absolute Configurations. *Chirality* **2022**, *34* (2), 333–363. https://doi.org/10.1002/chir.23393.

(282) Nicolaou, K. C.; Snyder, S. A. Chasing Molecules That Were Never There: Misassigned Natural Products and the Role of Chemical Synthesis in Modern Structure Elucidation. *Angew Chem Int Ed* **2005**, *44* (7), 1012–1044. https://doi.org/10.1002/anie.200460864.
(283) Chhetri, B. K.; Lavoie, S.; Sweeney-Jones, A. M.; Kubanek, J. Recent Trends in the Structural Revision of Natural Products. *Nat. Prod. Rep.* **2018**, *35* (6), 514–531. https://doi.org/10.1039/C8NP00011E.
(284) Menna, M.; Imperatore, C.; Mangoni, A.; Della Sala, G.; Taglialatela-Scafati, O. Challenges in the Configuration Assignment of Natural Products. A Case-Selective Perspective. *Nat. Prod. Rep.* **2019**, *36* (3), 476–489. https://doi.org/10.1039/C8NP00053K.
(285) Liu, H.-B.; Imler, G. H.; Baldridge, K. K.; O'Connor, R. D.; Siegel, J. S.; Deschamps, J. R.; Bewley, C. A. X-Ray Crystallography and Unexpected Chiroptical Properties Reassign the Configuration of Haliclonadiamine. *J. Am. Chem. Soc.* **2020**, *142* (6), 2755–2759. https://doi.org/10.1021/jacs.9b12926.
(286) Nakashima, Y.; Inoshita, T.; Kitajima, M.; Ishikawa, H. Asymmetric Total Synthesis of Senepodine F. *Org. Lett.* **2023**, *25* (7), 1151–1155. https://doi.org/10.1021/acs.orglett.3c00133.
(287) Zhang, Y.; Saha, S.; Esser, Y. C. C.; Ting, C. P. Total Synthesis and Stereochemical Assignment of Enteropeptin A. *J. Am. Chem. Soc.* **2024**, *146* (26), 17629–17635. https://doi.org/10.1021/jacs.4c06126.
(288) Irie, R.; Hitora, Y.; Watanabe, R.; Clark, H.; Suyama, Y.; Sekiya, S.; Suzuki, T.; Takada, K.; Matsunaga, S.; Hosokawa, S.; Oikawa, M. Stereochemical Assignment of the 36-Membered Macrolide Ring Portion of Poecillastrin C. *Org. Lett.* **2024**, *26* (25), 5290–5294. https://doi.org/10.1021/acs.orglett.4c01632.
(289) Kong, Y.; Liu, Y.; Wang, K.; Wang, T.; Wang, C.; Ai, B.; Jia, H.; Pan, G.; Yin, M.; Xu, Z. Confirmation of the Stereochemistry of Spiroviolene. *Beilstein J. Org. Chem.* **2024**, *20*, 852–858. https://doi.org/10.3762/bjoc.20.77.
(290) Shenvi, R. A. Natural Product Synthesis in the 21st Century: Beyond the Mountain Top. *ACS Cent. Sci.* **2024**, *10* (3), 519–528. https://doi.org/10.1021/acscentsci.3c01518.
(291) 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.
(292) Miyao, T.; Reker, D.; Schneider, P.; Funatsu, K.; Schneider, G. Chemography of Natural Product Space. *Planta med.* **2015**, DOI: 10.1055/s-0034-1396322. https://doi.org/10.1055/s-0034-1396322.
(293) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **2021**, *84* (11), 2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399.
(294) Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP Navigator: A New Look at the Natural Product Chemical Space. *Molecular Informatics* **2021**, *40* (9), 2100068. https://doi.org/10.1002/minf.202100068.
(295) Heinzke, A. L.; Pahl, A.; Zdrazil, B.; Leach, A. R.; Waldmann, H.; Young, R. J.; Leeson, P. D. Occurrence of "Natural Selection" in Successful Small Molecule Drug Discovery. *J. Med. Chem.* **2024**, *67* (13), 11226–11241. https://doi.org/10.1021/acs.jmedchem.4c00811.
(296) Jiang, J.; Ke, L.; Chen, L.; Dou, B.; Zhu, Y.; Liu, J.; Zhang, B.; Zhou, T.; Wei, G.-W. Transformer Technology in Molecular Science. *WIREs Computational Molecular Science* **2024**, *14* (4), e1725. https://doi.org/10.1002/wcms.1725.
(297) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer Learning Enables the Molecular Transformer to Predict Regio- and Stereoselective Reactions on Carbohydrates. *Nat. Commun.* **2020**, *11* (1), 4874. https://doi.org/10.1038/s41467-020-18671-7.
(298) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12* (25), 8648–8659. https://doi.org/10.1039/D1SC02362D.

(299) Kreutter, D.; Reymond, J.-L. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *Chem. Sci.* **2023**, *14* (36), 9959–9969. https://doi.org/10.1039/D3SC01604H.
(300) Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. arXiv 2017. https://doi.org/10.48550/ARXIV.1703.07076.
(301) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-Art Augmented NLP Transformer Models for Direct and Single-Step Retrosynthesis. *Nat Commun* **2020**, *11* (1), 5575. https://doi.org/10.1038/s41467-020-19266-y.
(302) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *J Cheminform* **2019**, *11* (1), 71. https://doi.org/10.1186/s13321-019-0393-0.
(303) Corrodi, H.; Hardegger, E. Herstellung des racemischen Colchicins und des unnatürlichen (+)-Colchicins. *Helv. Chim. Acta* **1957**, *40* (1), 193–199. https://doi.org/10.1002/hlca.19570400123.
(304) Wani, M. C.; Taylor, H. L.; Wall, M. E.; Coggon, P.; McPhail, A. T. Plant Antitumor Agents. VI. Isolation and Structure of Taxol, a Novel Antileukemic and Antitumor Agent from Taxus Brevifolia. *J. Am. Chem. Soc.* **1971**, *93* (9), 2325–2327. https://doi.org/10.1021/ja00738a045.
(305) 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.
(306) Pettit, G. R.; Singh, S. B.; Hogan, F.; Burkett, D. D. Chiral Modifications of Dolastatin 10: The Potent Cytostatic Peptide (19aR)-Isodolastatin 10. *J. Med. Chem.* **1990**, *33* (12), 3132–3133. https://doi.org/10.1021/jm00174a006.
(307) Butenandt, A.; Beckmann, R.; Hecker, E. Über Den Sexuallockstoff Des Seidenspinners, I. Der Biologische Test Und Die Isolierung Des Reinen Sexuallockstoffes Bombykol. *Biol. Chem.* **1961**, *324*, 71–83. https://doi.org/10.1515/bchm2.1961.324.1.71.
(308) Staudinger, H.; Ruzicka, L. Insektentötende Stoffe III. Konstitution Des Pyrethrolons. *Helv. Chim. Acta* **1924**, *7* (1), 212–235. https://doi.org/10.1002/hlca.19240070126.
(309) Manosroi, A.; Jantrawut, P.; Ogihara, E.; Yamamoto, A.; Fukatsu, M.; Yasukawa, K.; Tokuda, H.; Suzuki, N.; Manosroi, J.; Akihisa, T. Biological Activities of Phenolic Compounds and Triterpenoids from the Galls of Terminalia Chebula. *Chem. Biodiv.* **2013**, *10* (8), 1448–1463. https://doi.org/10.1002/cbdv.201300149.
(310) Mijts, B. N.; Lee, P. C.; Schmidt-Dannert, C. Identification of a Carotenoid Oxygenase Synthesizing Acyclic Xanthophylls: Combinatorial Biosynthesis and Directed Evolution. *Chemistry & Biology* **2005**, *12* (4), 453–460. https://doi.org/10.1016/j.chembiol.2005.02.010.
(311) Mayer, H.; Schudel, P.; Rüegg, R.; Isler, O. Über Die Chemie Des Vitamins E. 3. Mitteilung. Die Totalsynthese von (2R, 4′R, 8′R)- Und (2S, 4′R, 8′R)-α-Tocopherol. *Helv. Chim. Acta* **1963**, *46* (2), 650–671. https://doi.org/10.1002/hlca.19630460225.
(312) Zapesochnaya, G. G.; Tyukavkina, N. A.; Eremin, S. K. Flavonoids ofDatisca Cannibina. VI. Properties of Datiscin. *Chem. Nat. Compd.* **1982**, *18* (2), 163–166. https://doi.org/10.1007/BF00577184.
(313) Tu, L.; Shen, S.; Yan, Z.; Li, X.; Liu, K.; Xu, J.; Luo, M. Discovery of Olimycin E from Streptomyces Sp. 11695. *Nat. Prod. Res.* **2024**, *0* (0), 1–6. https://doi.org/10.1080/14786419.2024.2337131.
(314) Koseki, Y.; Nishimura, H.; Asano, R.; Aoki, K.; Shiyu, L.; Sugiyama, R.; Yamazaki, M. Isolation of New Indole Alkaloid Triglucoside from the Aqueous Extract of Uncaria Rhynchophylla. *J. Nat. Med.* **2025**, *79* (1), 28–35. https://doi.org/10.1007/s11418-024-01836-9.
(315) Yang, J.; Yao, F.-H.; Xu, S.-F.; Shi, J.-Y.; Li, X.-Y.; Yi, X.-X.; Gao, C.-H. Mauritone A, a New Polyketide from a Fungal-Bacterial Symbiont Aspergillus Spelaeus GXIMD

04541/Sphingomonas Echinoides GXIMD 04532. *Nat. Prod. Res.* **2024**, *0* (0), 1–6. https://doi.org/10.1080/14786419.2024.2377313.
(316) Batista, A. N. L.; Santos, C. H. T.; de Albuquerque, A. C. F.; Santos Jr., F. M.; Garcez, F. R.; Batista Jr., J. M. Absolute Configuration Reassignment of Nectamazin A: Implications to Related Neolignans. *Spectrochim. Acta A: Mol. Biomol. Spectr.* **2024**, *304*, 123283. https://doi.org/10.1016/j.saa.2023.123283.
(317) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*; Bansal, M., Ji, H., Eds.; Association for Computational Linguistics: Vancouver, Canada, 2017; pp 67–72.

Appendix

Appendix A - Supplementary information for: Alchemical Analysis of FDA Approved Drugs

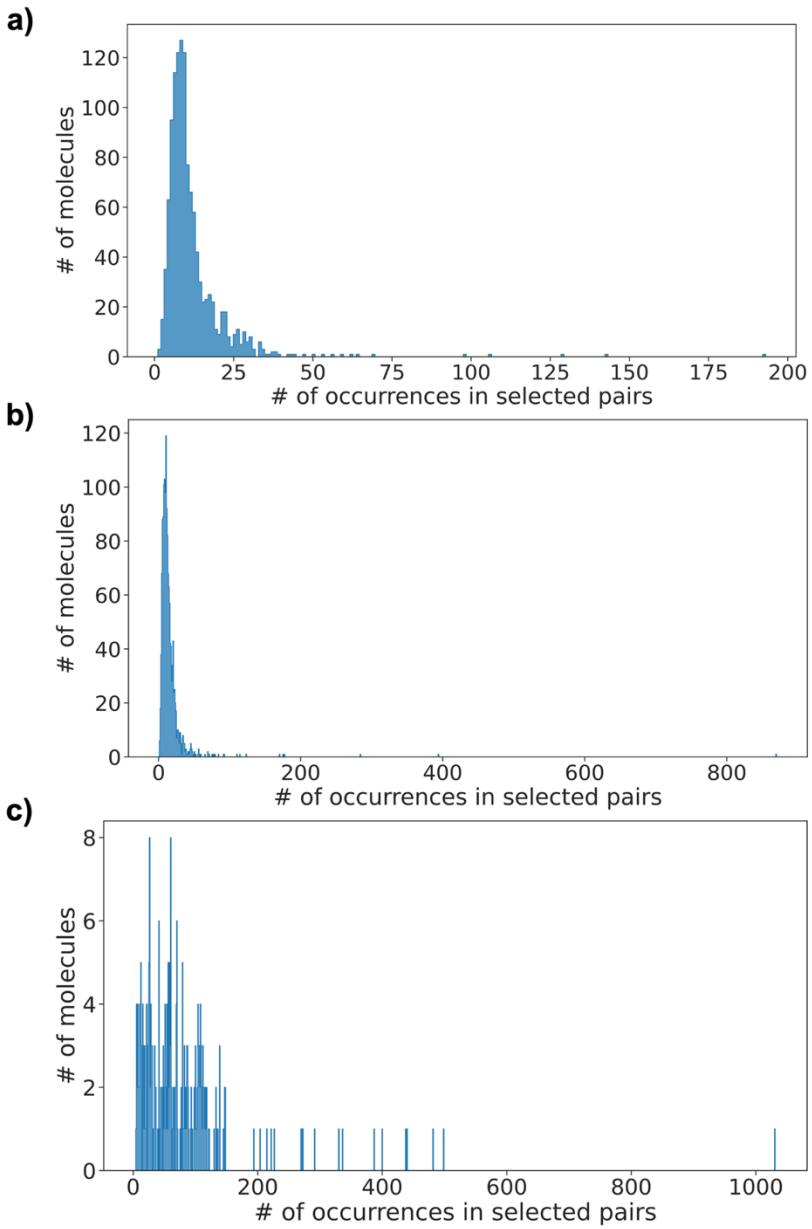


Figure A1. Count of molecules by number of occurrences in selected pairs for the a) FDA, b) EGFR and c) PMB set. In all sets, most of the molecules appear sporadically in the selected pairs. Only a limited number of compounds appears more often**.**

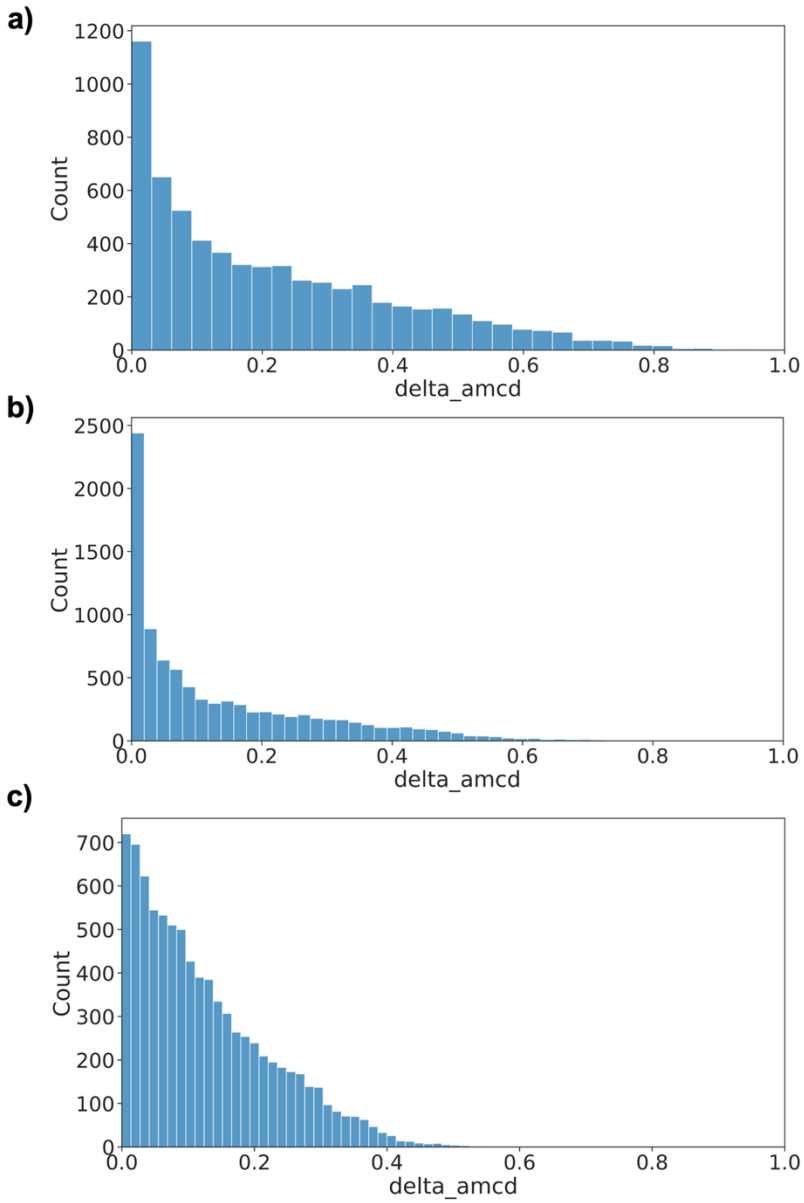


Figure A2. Count of molecular pairs by difference in atom-mapping confidence score between the forward and backward reactions in selected pairs for the a) FDA, b) EGFR and c) PMB set.

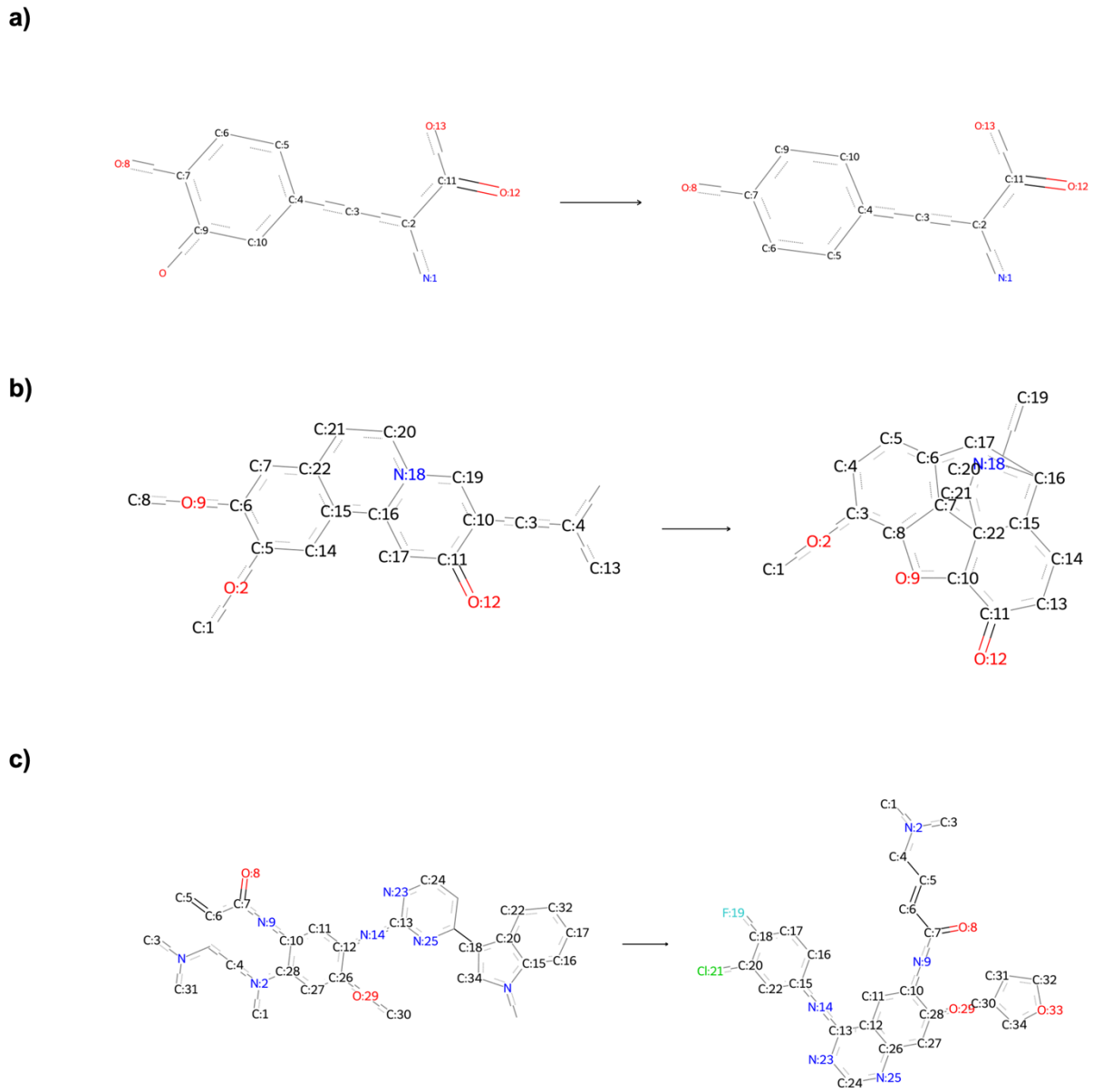


Figure A3. Full atom mapping of the examples selected from the FDA set. The shown atom-mapping is the one of the backwards reactions.

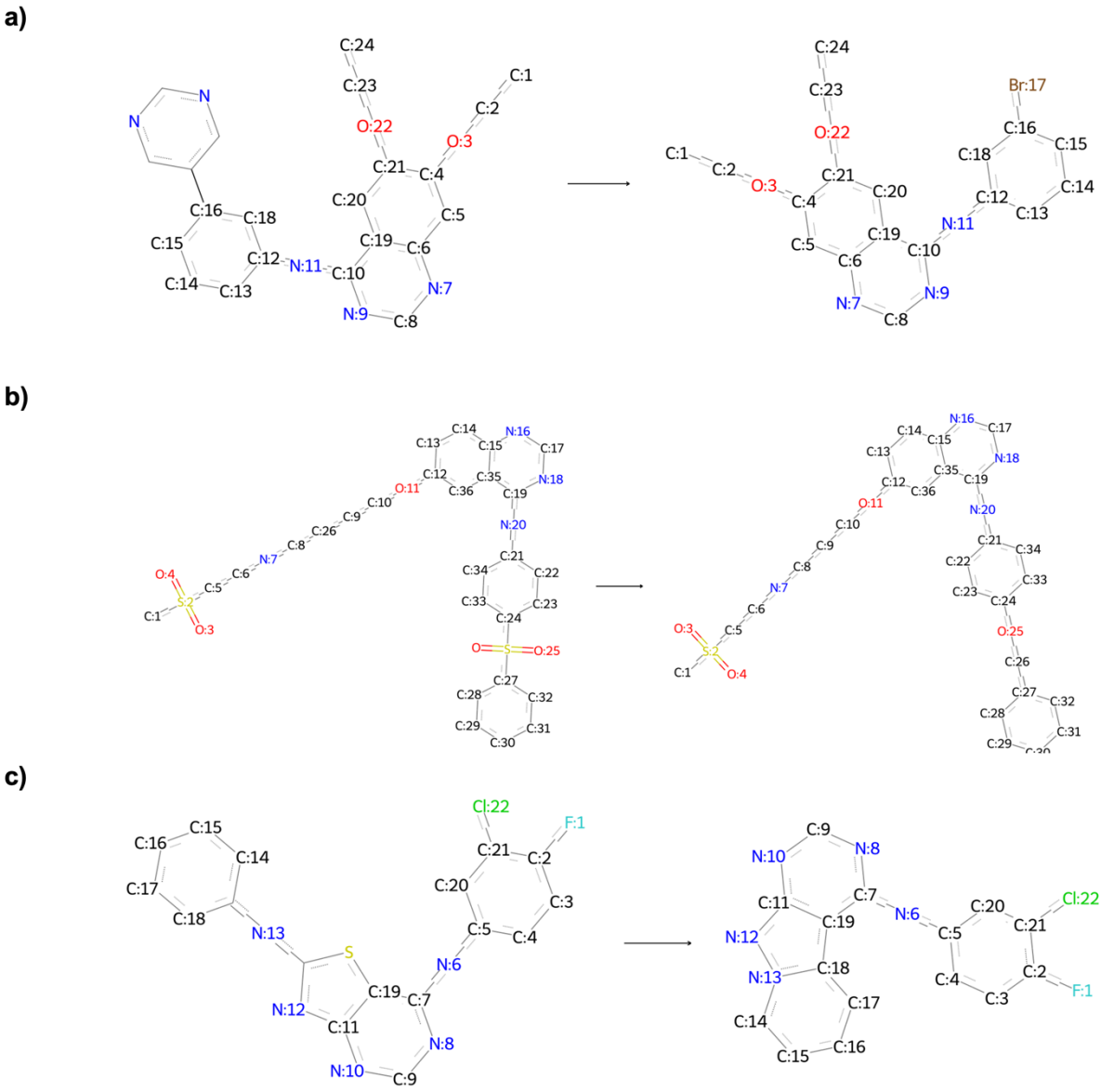


Figure A4. Full atom mapping of the examples selected from the EGFR set. The shown atom-mapping is the one of the backwards reactions.

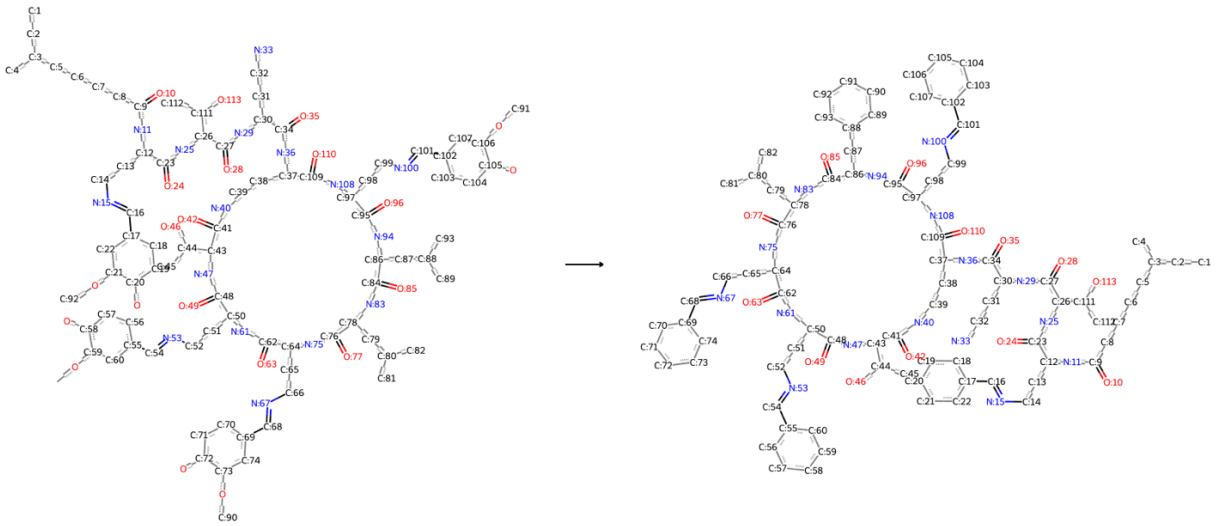


Figure A5. Full atom mapping of the example selected from the PMB set. The shown atom-mapping is the one of the backwards reaction.

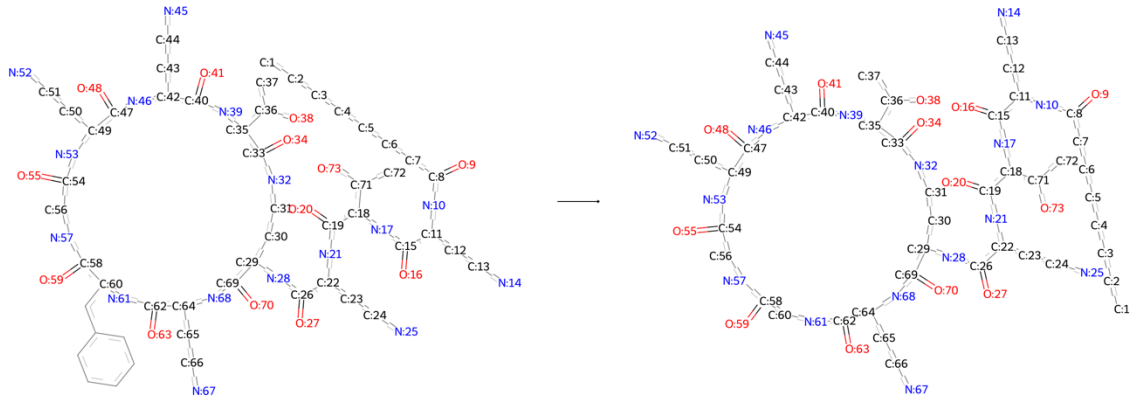


Figure A6. Full atom-mapping of mutation of a glycine to a phenylalanine residue (amcd: 0.32), corresponding to a feasible α-alkylation reaction of glycine with benzyl bromide.

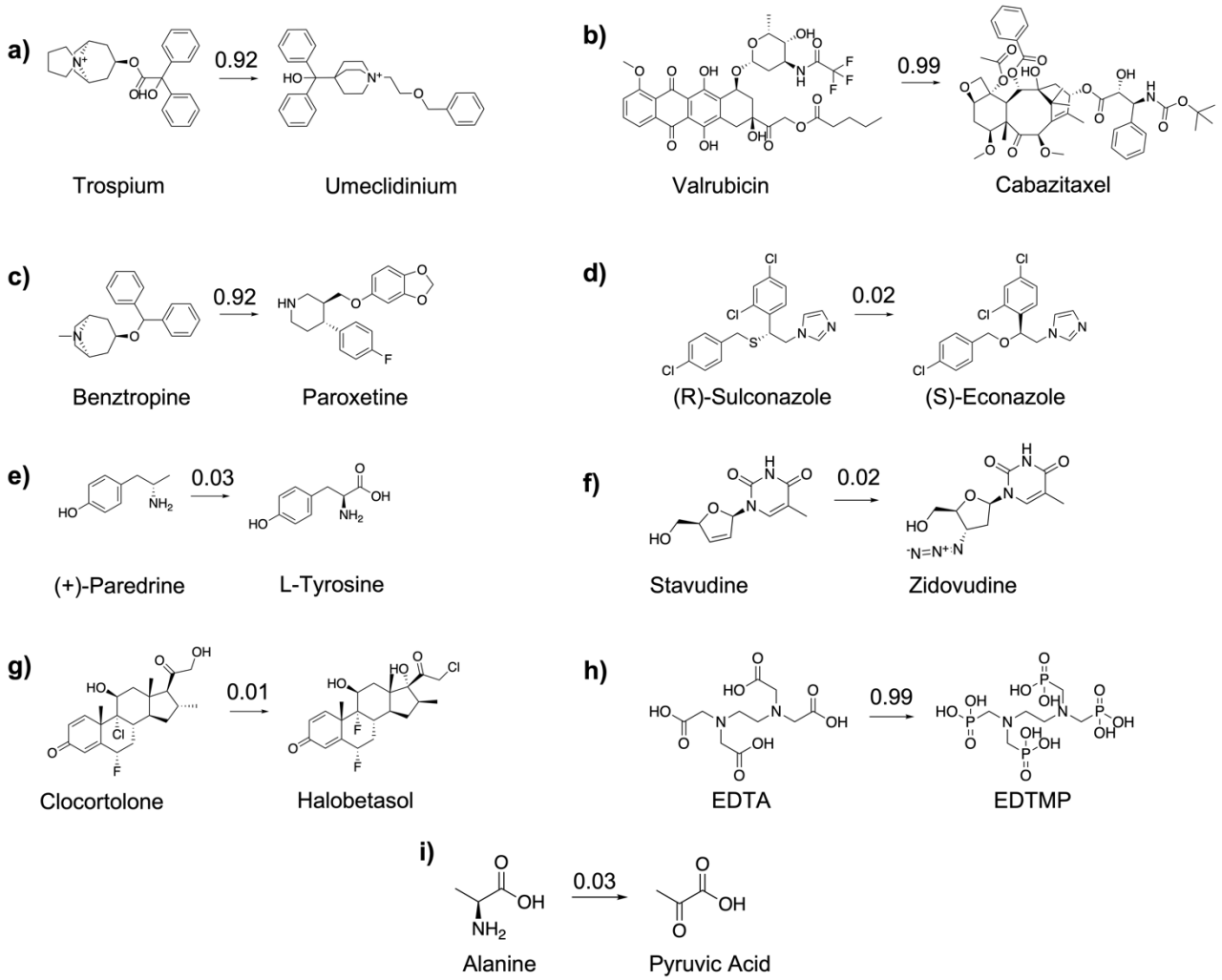


Figure A7. Additional interesting pairs selected from the drug pairs in the FDA-approved subset and the determined atom-mapping confidence distance of the reaction. a) Trospium and Umeclidinium, two anticholinergic drugs acting on the muscarinic receptor. The structures contain common elements, such as the diphenylmethanol and tropane-like moieties, which are completely rearranged between the two structures. b) Valrubicin and Cabazitaxel, two anticancer drugs acting on topoisomerase II and tubulin stabilization respectively. Although the two compounds act on different targets, these targets are part of the same pathway and their inhibition leads to cell death. c) Benztropine and Paroxetine, two unrelated drugs acting on serotonin uptake inhibition. d) (R)-Sulconazole and (S)-Econazole, two imidazole antifungals differing from each other by a single atom mutation from S to O. e) (+)-Paredrine and L-Tyrosine, two closely related structures separated by an alchemical condensation of a carboxylic acid to a methyl and stereo-inversion. f) Stavudine and Zidovudine, two HIV reverse transcriptase inhibitors separated by an azidation. g) Clocortolone and Halobetasol, two steroid drugs used for the treatment of inflammatory and itching skin diseases. h) EDTA and EDTMP, both highly related chelating agents. i) Alanine and Pyruvic Acid, two highly related compounds separated by a N to =O mutation.

Appendix B - Supplementary information for: One chiral fingerprint to find them all

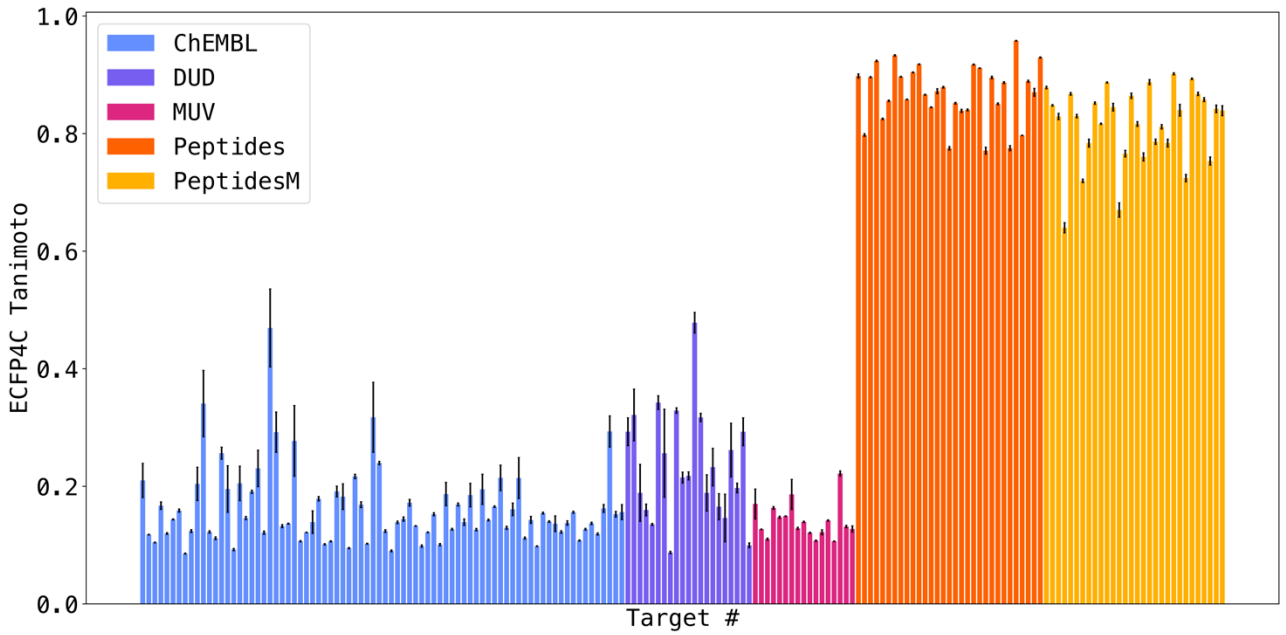


Figure B1. Mean and standard deviation of the pairwise similarities calculated for all 5 selected actives of each dataset contained in the benchmarking platform. Actives are encoded using the chiral ECFP4 (radius=2, nBits=2048) fingerprint and Tanimoto similarities determined for all possible pairs. ChEMBL, DUD and MUV sets comprise the original Riniker & Landrum benchmark. The "Peptides" set contains scrambled sequences of the same peptide. The "PeptidesM" set contains single point mutants of the same peptide.

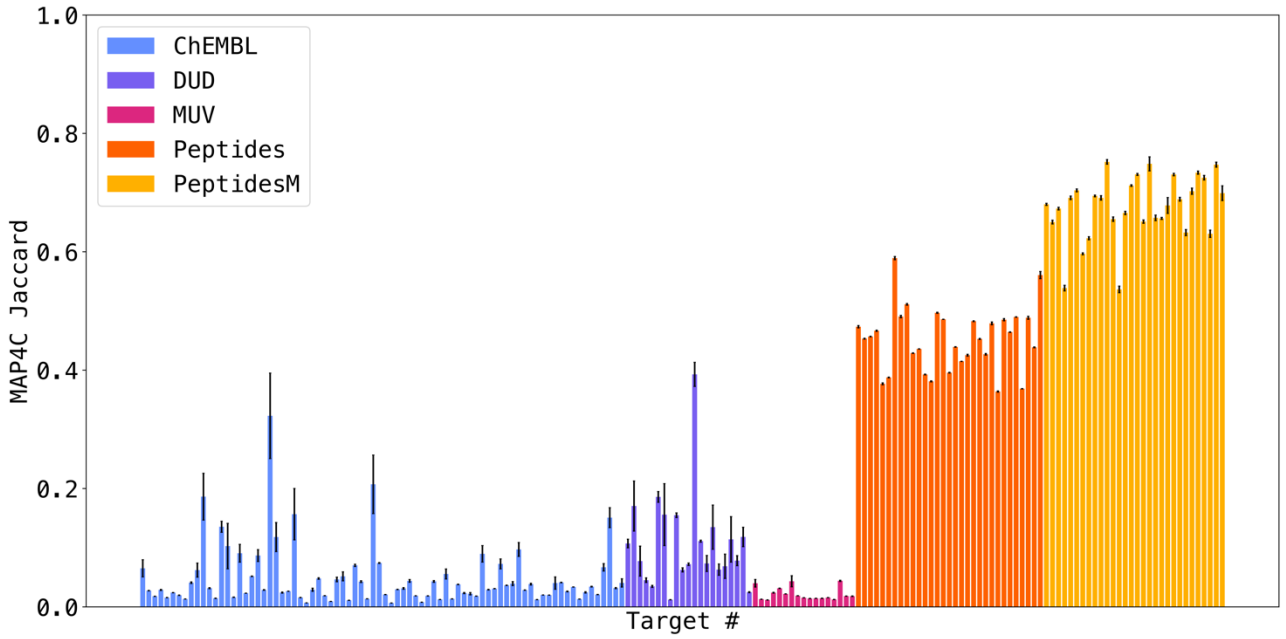


Figure B2. Mean and standard deviation of the pairwise similarities calculated for all 5 selected actives of each dataset contained in the benchmarking platform. Actives are encoded using the MAP4C (max_radius=2, n_permutations=2048) fingerprint and Jaccard similarities determined for all possible pairs. ChEMBL, DUD and MUV sets comprise the original Riniker & Landrum benchmark. The "Peptides" set contains scrambled sequences of the same peptide. The "PeptidesM" set contains single point mutants of the same peptide.

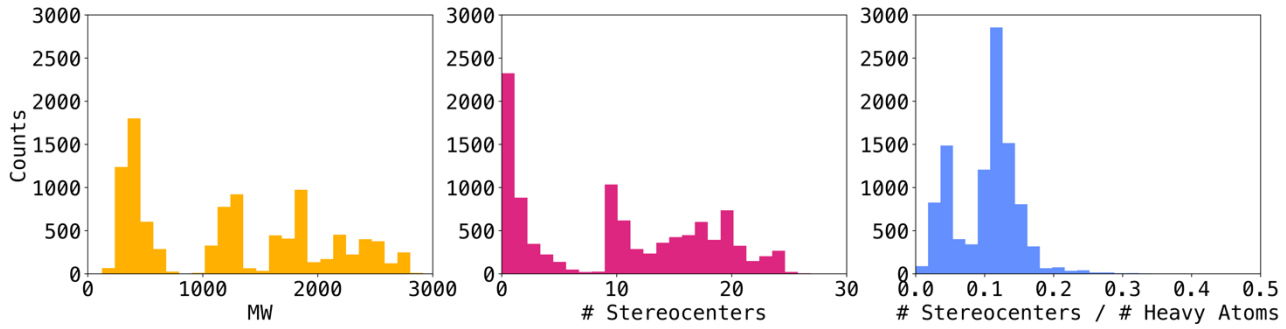


Figure B3. Distribution of molecular weight (MW) (yellow), number of stereocenters (magenta) and ratio of stereocenters to heavy atom count (blue) in the set uniformly sampled from the extended benchmark. The set contained a total of 10,122 compounds and was used to determine the relative impact of stereochemistry encoding on total similarity.

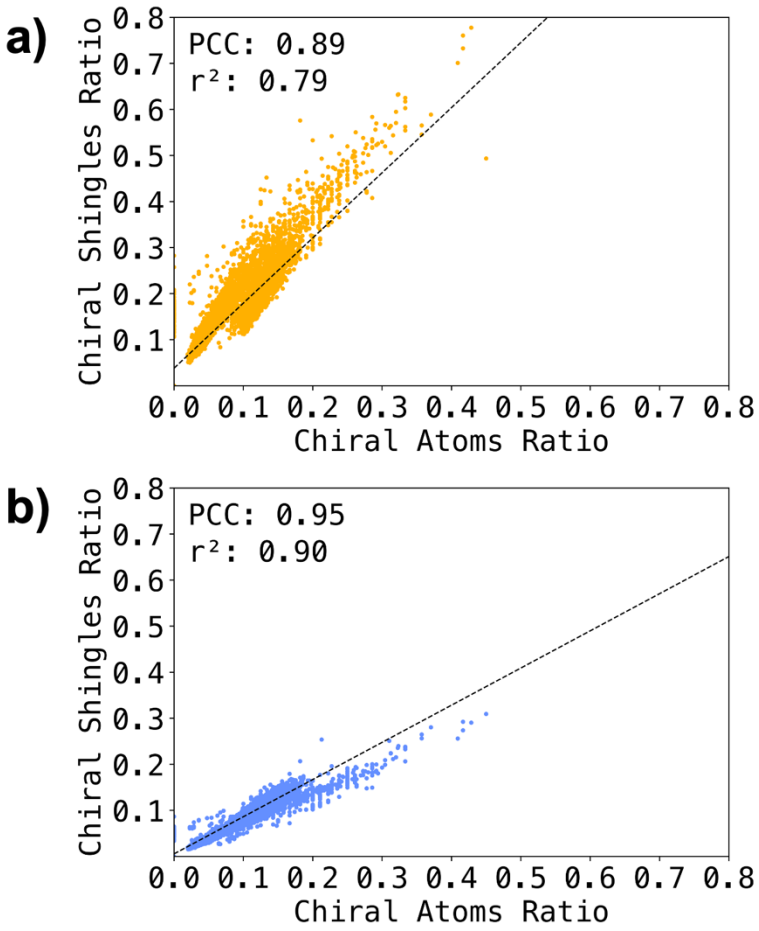


Figure B4. Scatterplots of chiral shingle ratio vs. chiral atoms ratio for a) radius = 1 b) radius = 2 and c) radius = 3. Additionally, the r^2 of the linear fit and the Pearson correlation coefficient (PCC) are reported. All reported PCCs are statistically significant.

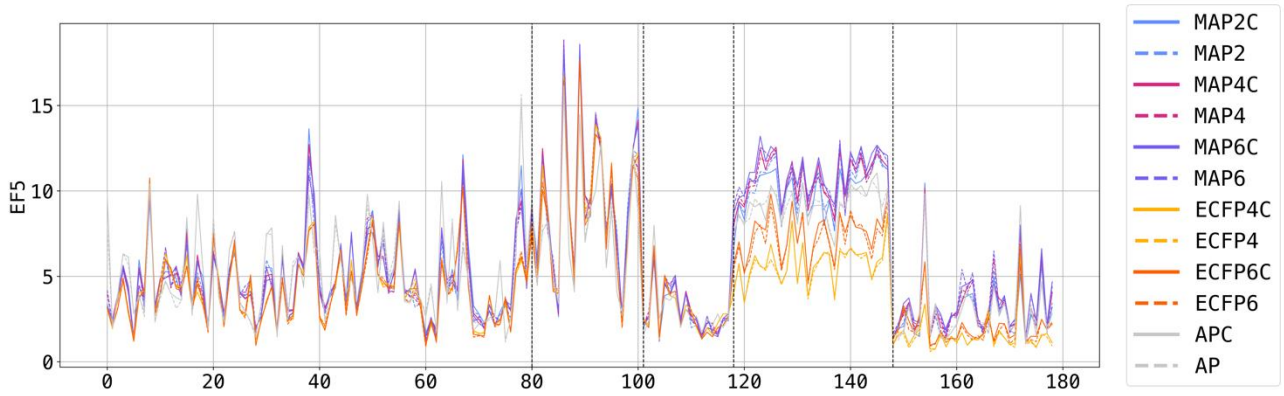


Figure B5. EF5 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

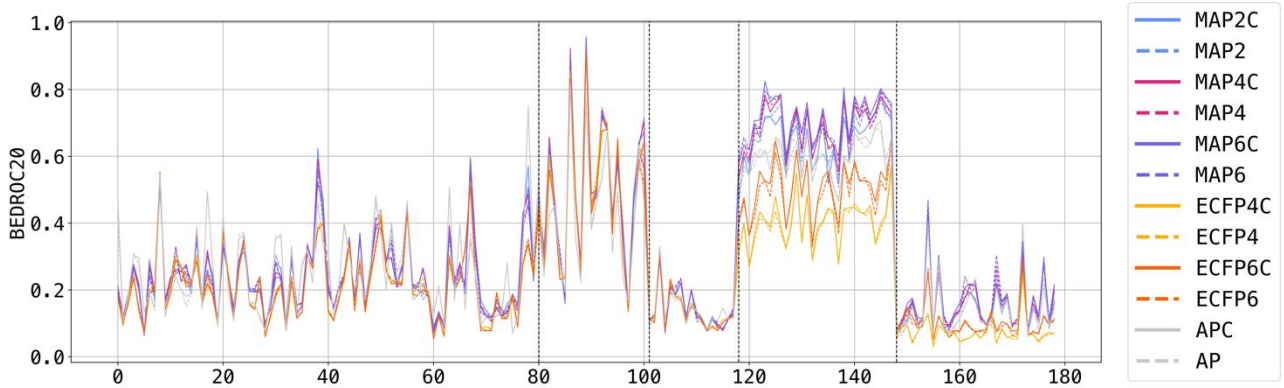


Figure B6. BEDROC20 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

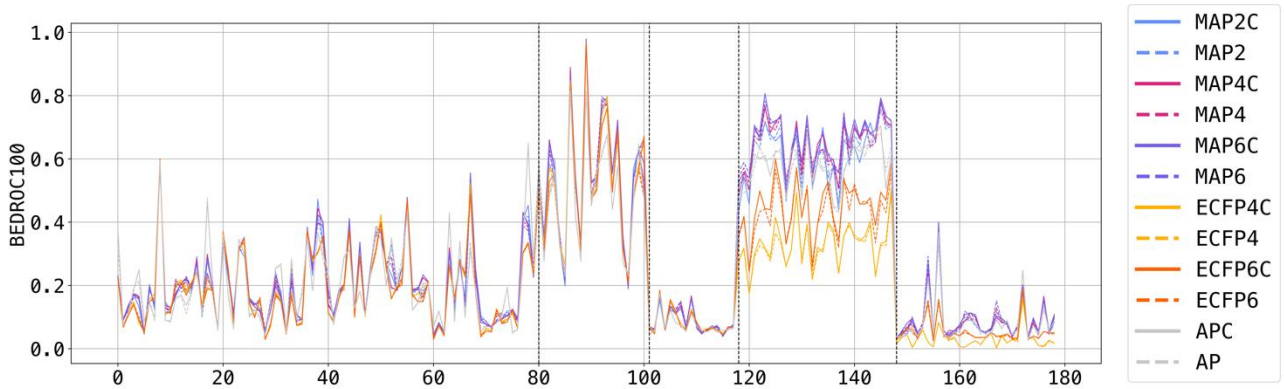


Figure B7. BEDROC100 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

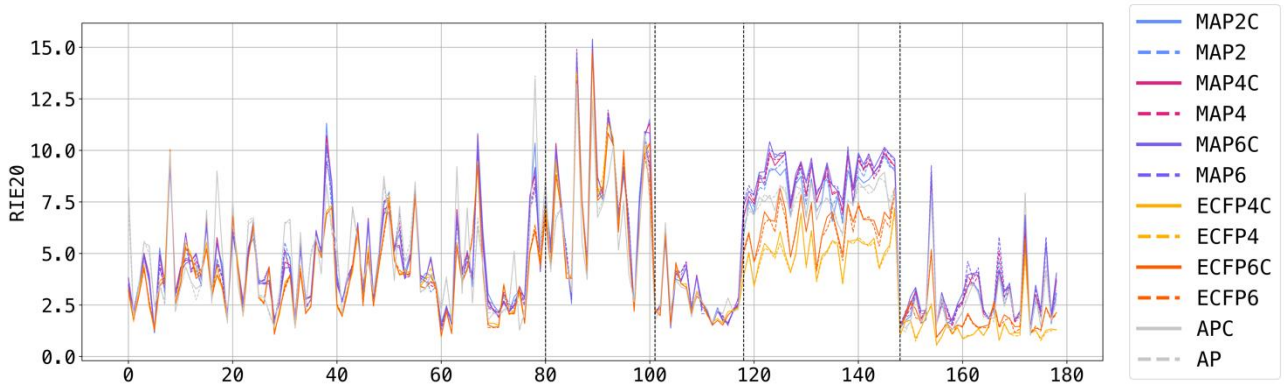


Figure B8. RIE20 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

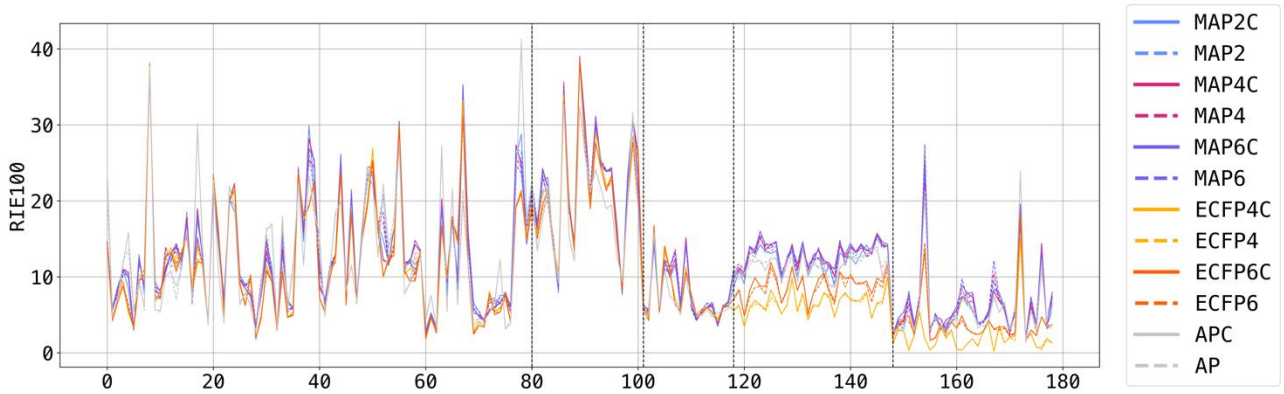


Figure B9. RIE100 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

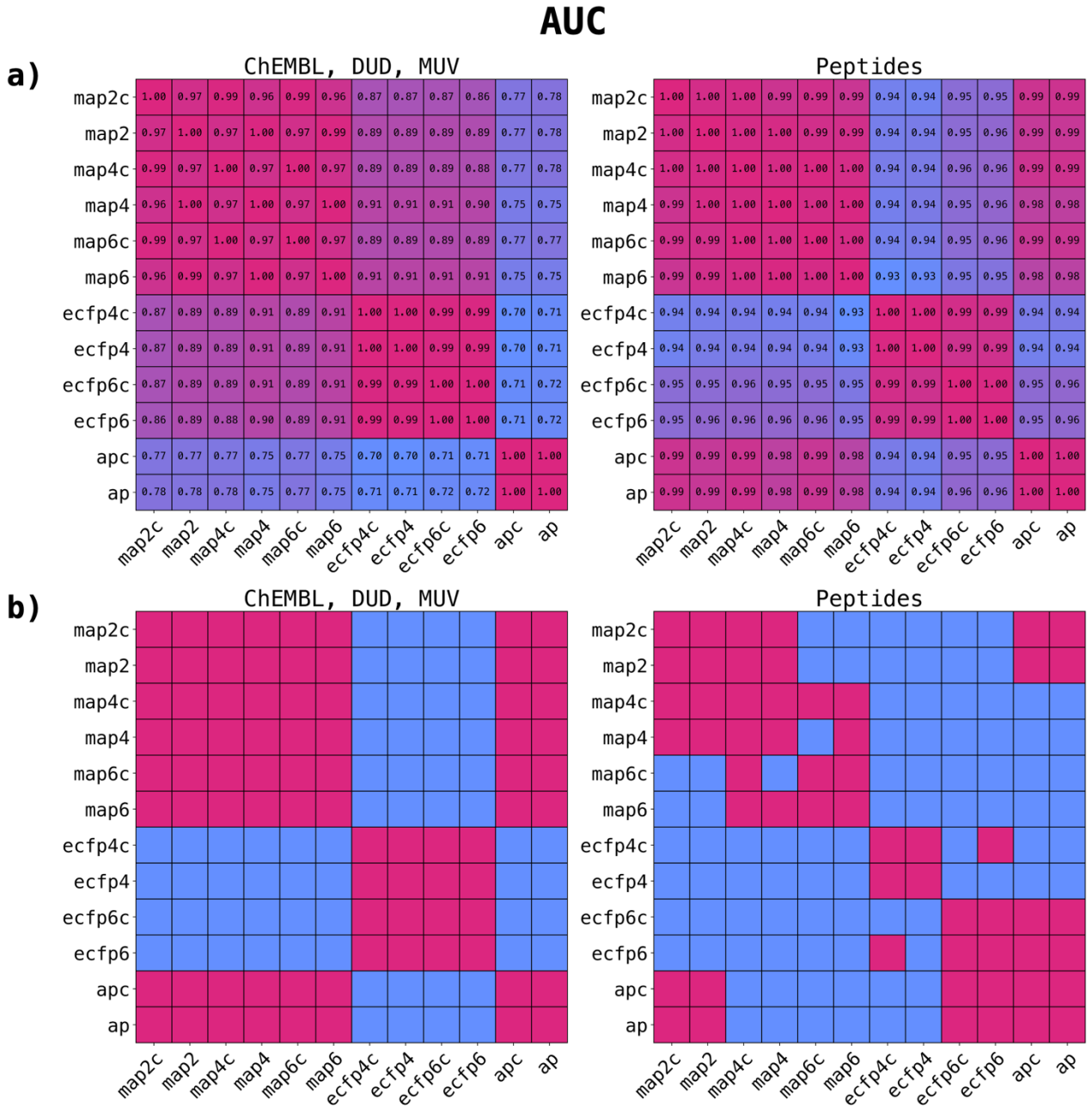


Figure B10. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean AUCs acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked AUCs from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

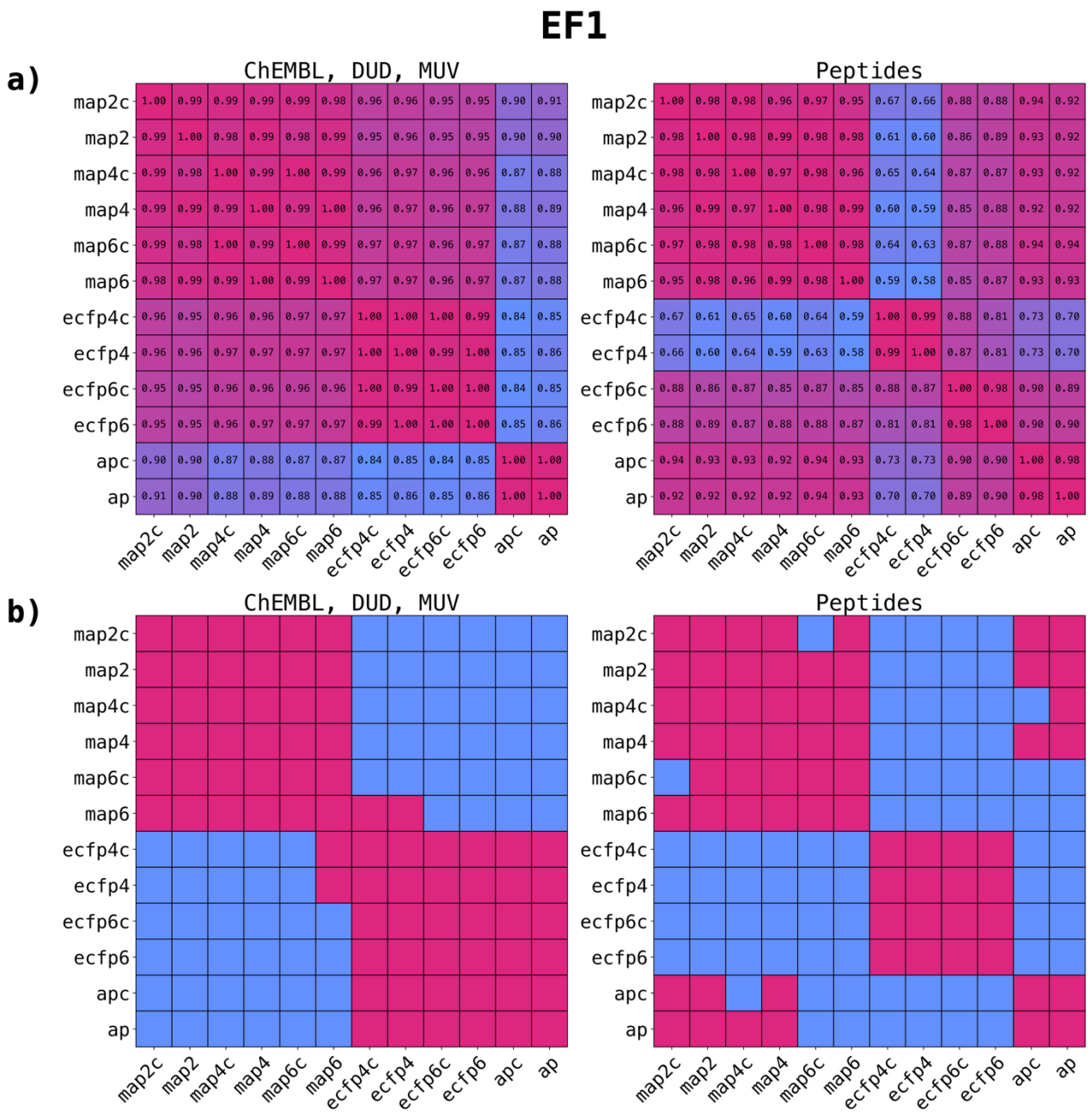


Figure B11. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean EF1s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked EF1s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

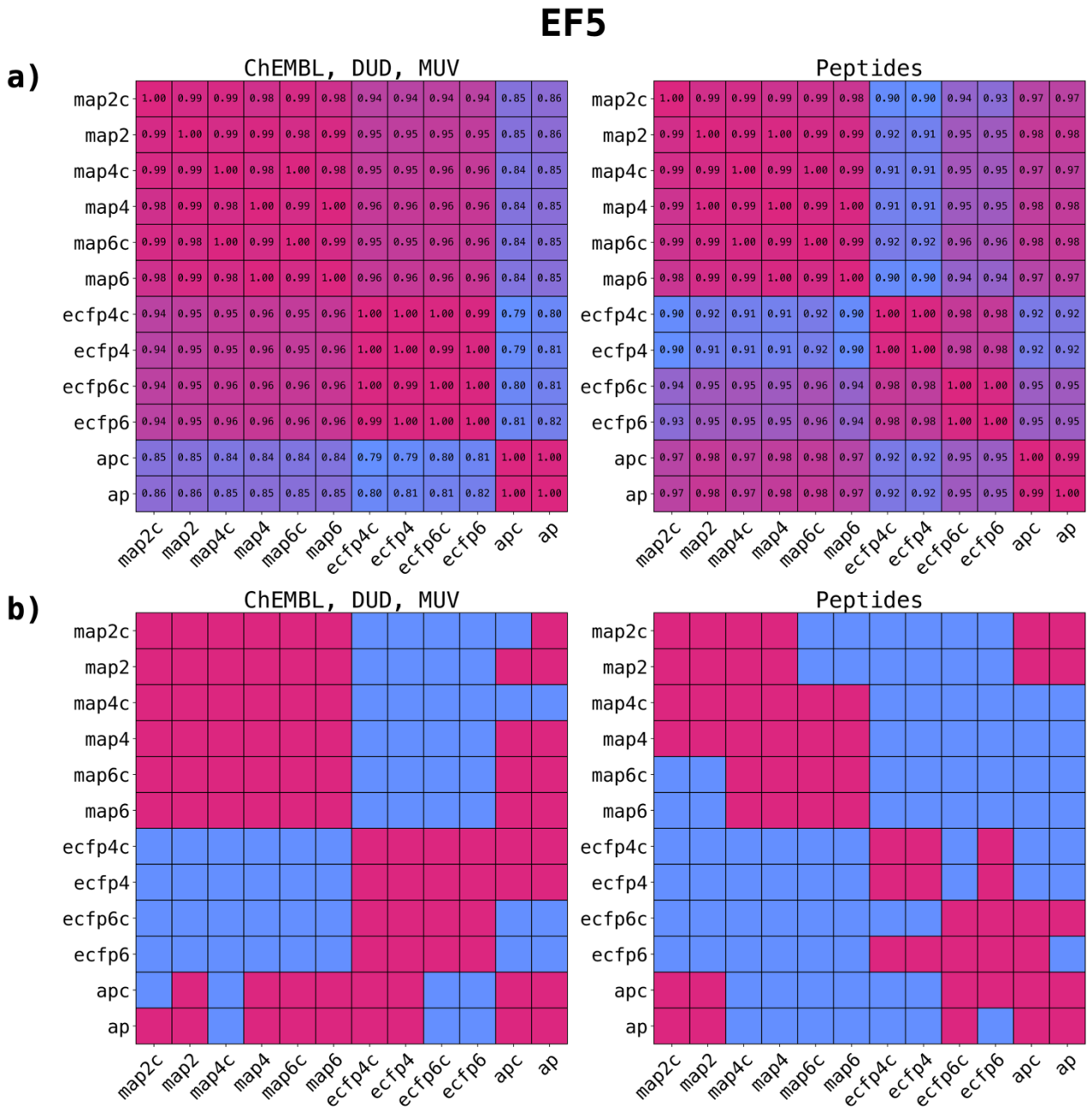


Figure B12. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean EF5s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked EF5s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

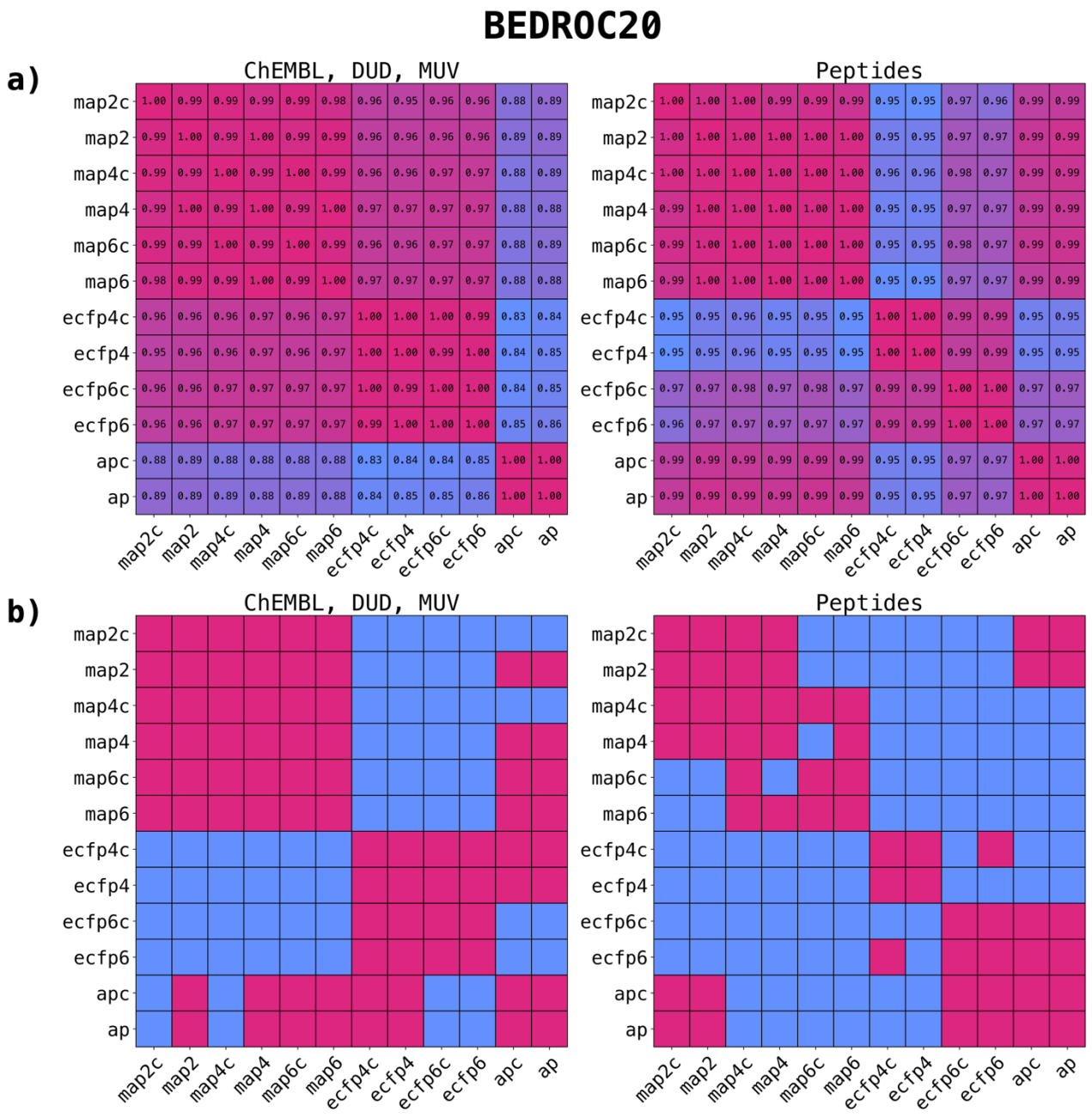


Figure B13: a) Pairwise Pearson correlations among tested fingerprints, derived from the mean BEDROC20s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked BEDROC20s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

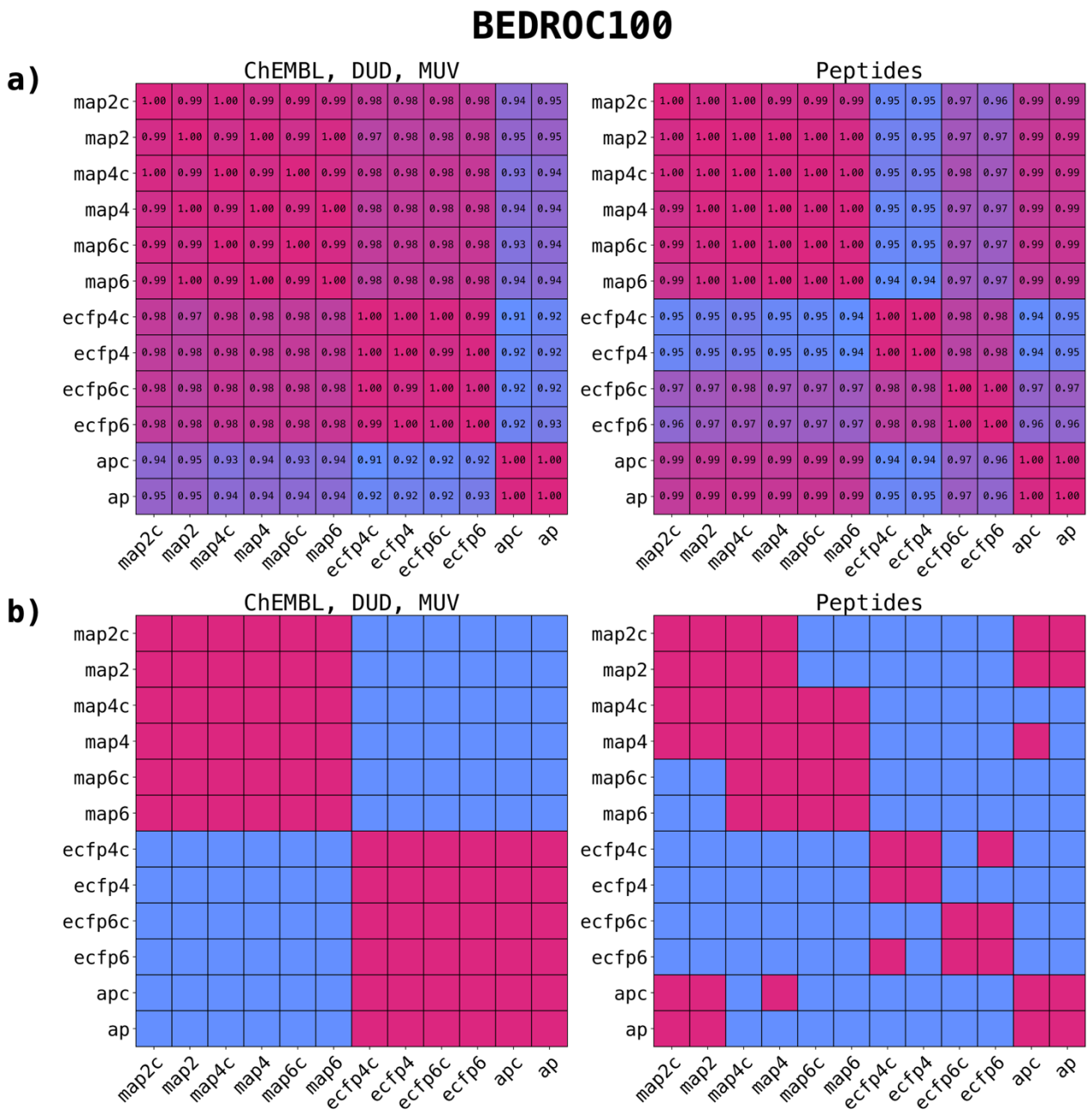


Figure B14. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean BEDROC100s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked BEDROC100s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

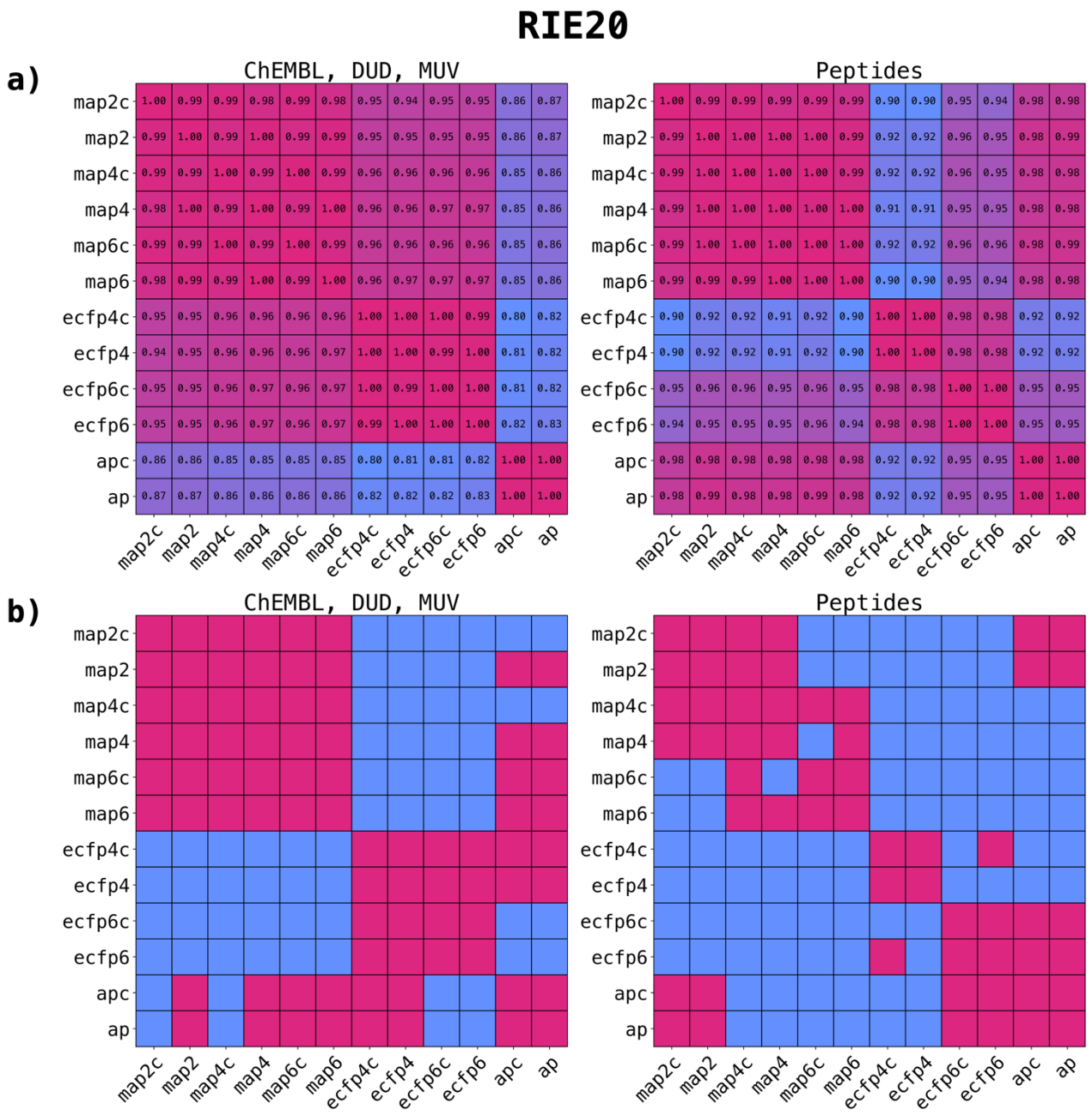


Figure B15. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean RIE20s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked RIE20s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

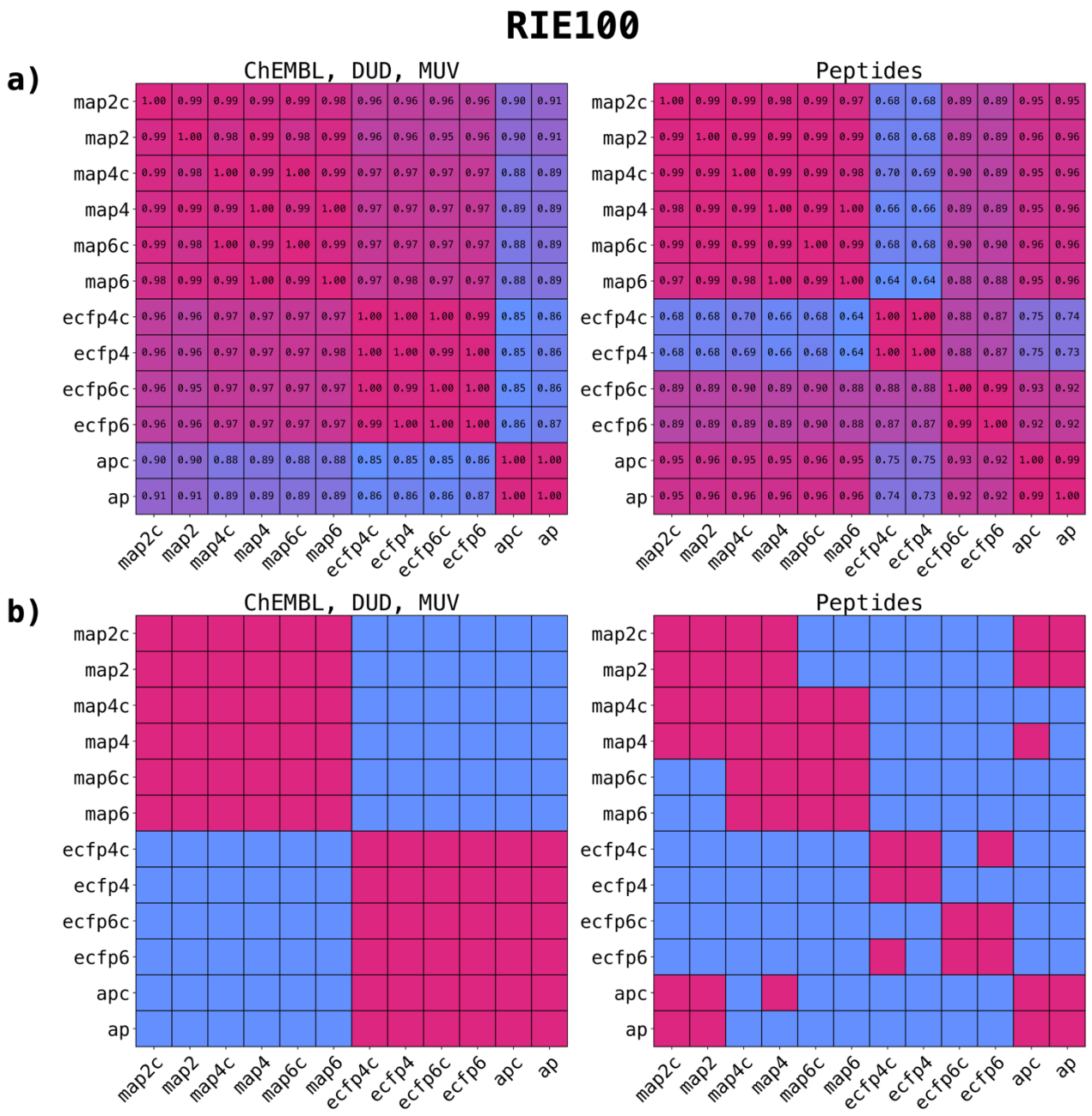


Figure B16. a) Pairwise Pearson correlations among tested fingerprints, derived from the mean RIE100s acquired from benchmark datasets. The numbers represent the Pearson correlation coefficient for each pair. b) Pairwise Friedman-Nemenyi test among tested fingerprints, based on the ranked RIE100s from benchmark datasets. A red square denotes a not significant difference between fingerprints at α=0.05, while a blue square denotes a significant difference.

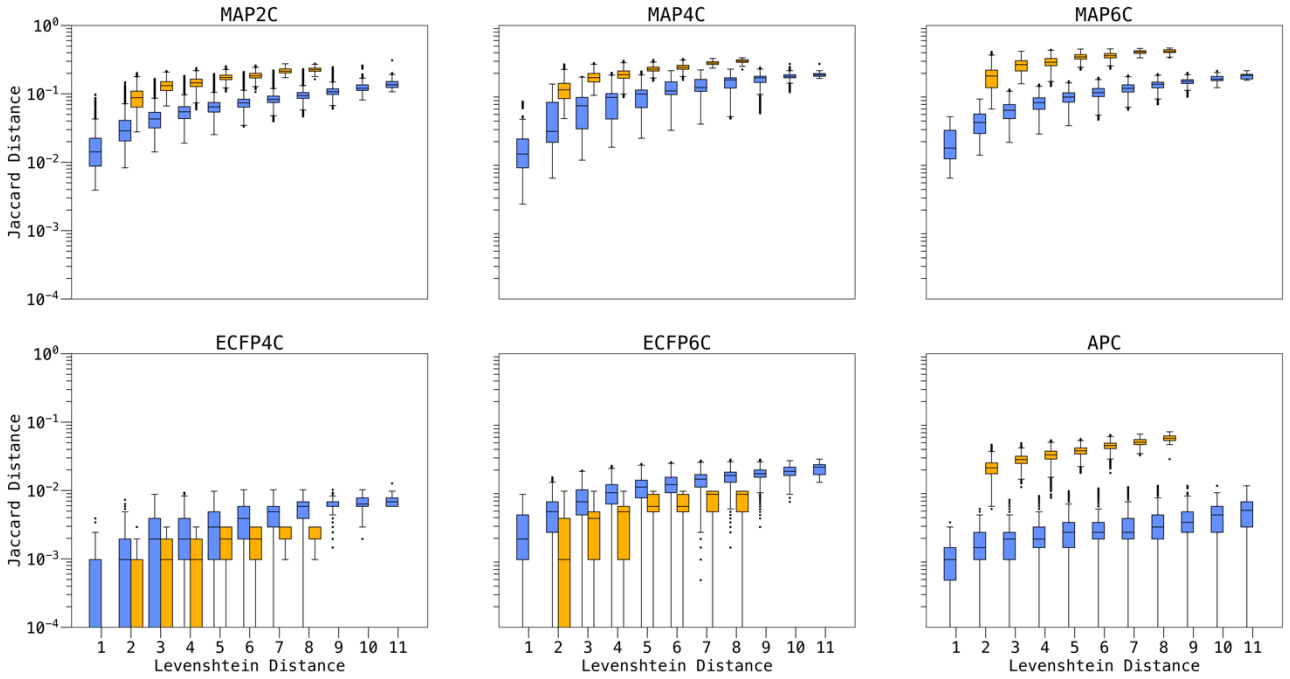


Figure B17. Comparative analysis of MAP2C, MAP4C, MAP6C, APC, ECFP4C and ECFP6C Jaccard distance assignment on ln65 diastereomers (blue) and structural isomers (yellow). The distance distributions are grouped by Levenshtein distance, used to determine the number of mutations from any sequence to ln65. MAPC fingerprints display a higher performance than the other fingerprints when it comes to distinguishing all possible diastereomers and structural isomers from each other. This is not the case for APC, which has difficulties distinguishing diastereomers, and ECPFC fingerprints, which cannot distinguish diastereomers or structural isomers robustly. MAPC fingerprints also consistently assign lower distances to diastereomers than structural isomers. APC follows the same trend, although the lower diastereomer distances are skewed due to the APC fingerprint not being able to robustly distinguish all diastereomers. ECFPC show a complete overlap of Jaccard distances for diastereomers and structural isomers. Finally, the overall Jaccard distances increase with increasing Levenshtein distance for MAPC fingerprints, indicating that the obtained distances align with intuitive changes such as stereocenter or residue mutations.

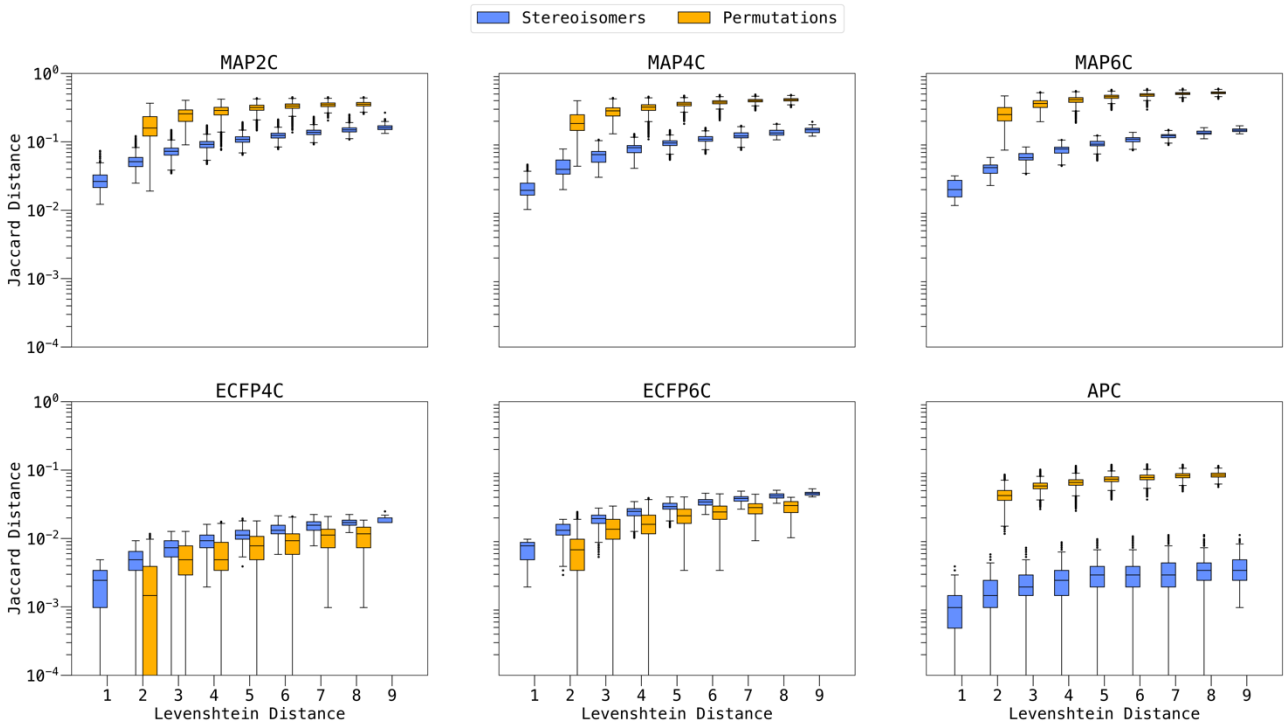
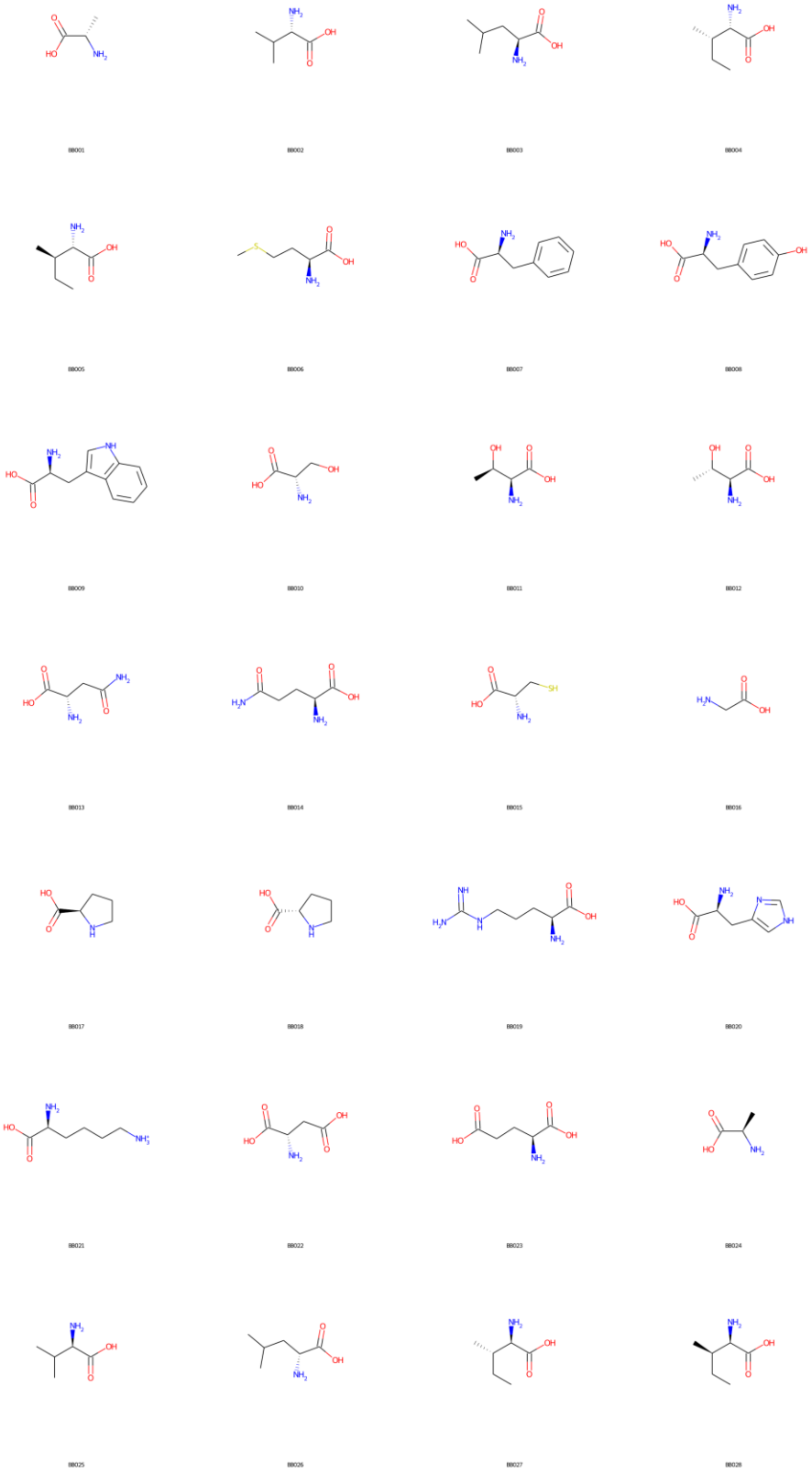
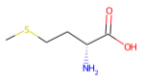


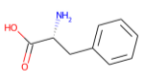
Figure B18. Comparative analysis of MAP2C, MAP4C, MAP6C, APC, ECFP4C and ECFP6C Jaccard distance assignment on polymyxin B2 diastereomers (blue) and structural isomers (yellow). The distance distributions are grouped by Levenshtein distance, used to determine the number of mutations from any sequence to polymyxin B2. MAPC fingerprints display a higher performance than the other fingerprints when it comes to distinguishing all possible diastereomers and structural isomers from each other. This is not the case for APC, which has difficulties distinguishing diastereomers, and ECPFC fingerprints, which cannot distinguish diastereomers or structural isomers robustly. MAPC fingerprints also consistently assign lower distances to diastereomers than structural isomers. APC follows the same trend, although the lower diastereomer distances are skewed due to the APC fingerprint not being able to robustly distinguish all diastereomers. ECFPC show a complete overlap of Jaccard distances for diastereomers and structural isomers. Finally, the overall Jaccard distances increase with increasing Levenshtein distance for MAPC fingerprints, indicating that the obtained distances align with intuitive changes such as stereocenter or residue mutations.

Appendix C - Supplementary information for: Navigating a 10E+ 60 chemical space of peptide/peptoid oligomers

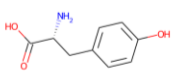




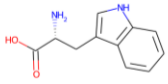
BB029



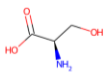
BB030



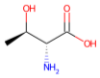
BB031



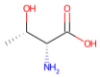
BB032



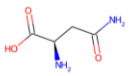
BB033



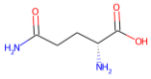
BB034



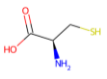
BB035



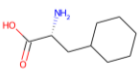
BB036



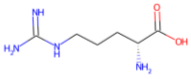
BB037



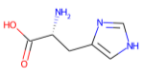
BB038



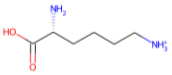
BB039



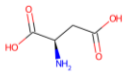
BB040



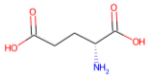
BB041



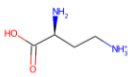
BB042



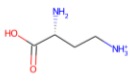
BB043



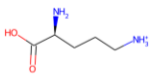
BB044



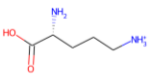
BB045



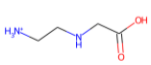
BB046



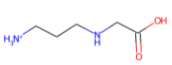
BB047



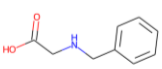
BB048



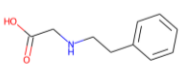
BB049



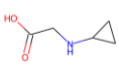
BB050



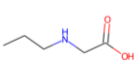
BB051



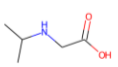
BB052



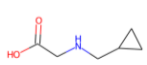
BB053



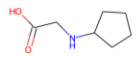
BB054



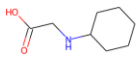
BB055



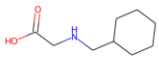
BB056



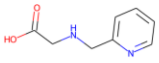
BB057



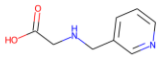
BB058



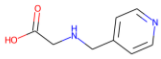
BB059



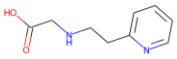
BB060



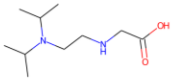
BB061



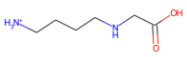
BB062



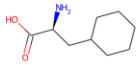
BB063



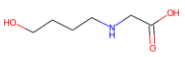
BB064



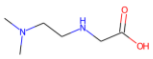
BB065



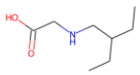
BB066



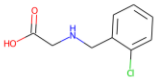
BB067



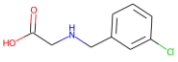
BB068



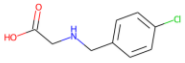
BB069



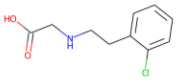
BB070



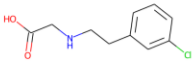
BB071



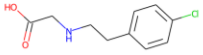
BB072



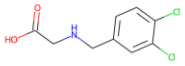
BB073



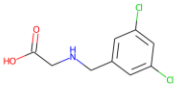
BB074



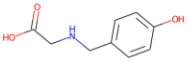
BB075



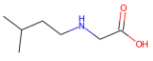
BB076



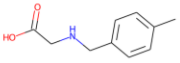
BB077



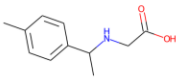
BB078



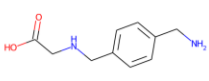
BB079



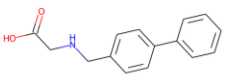
BB080



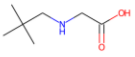
BB081



BB082



BB083



BB084

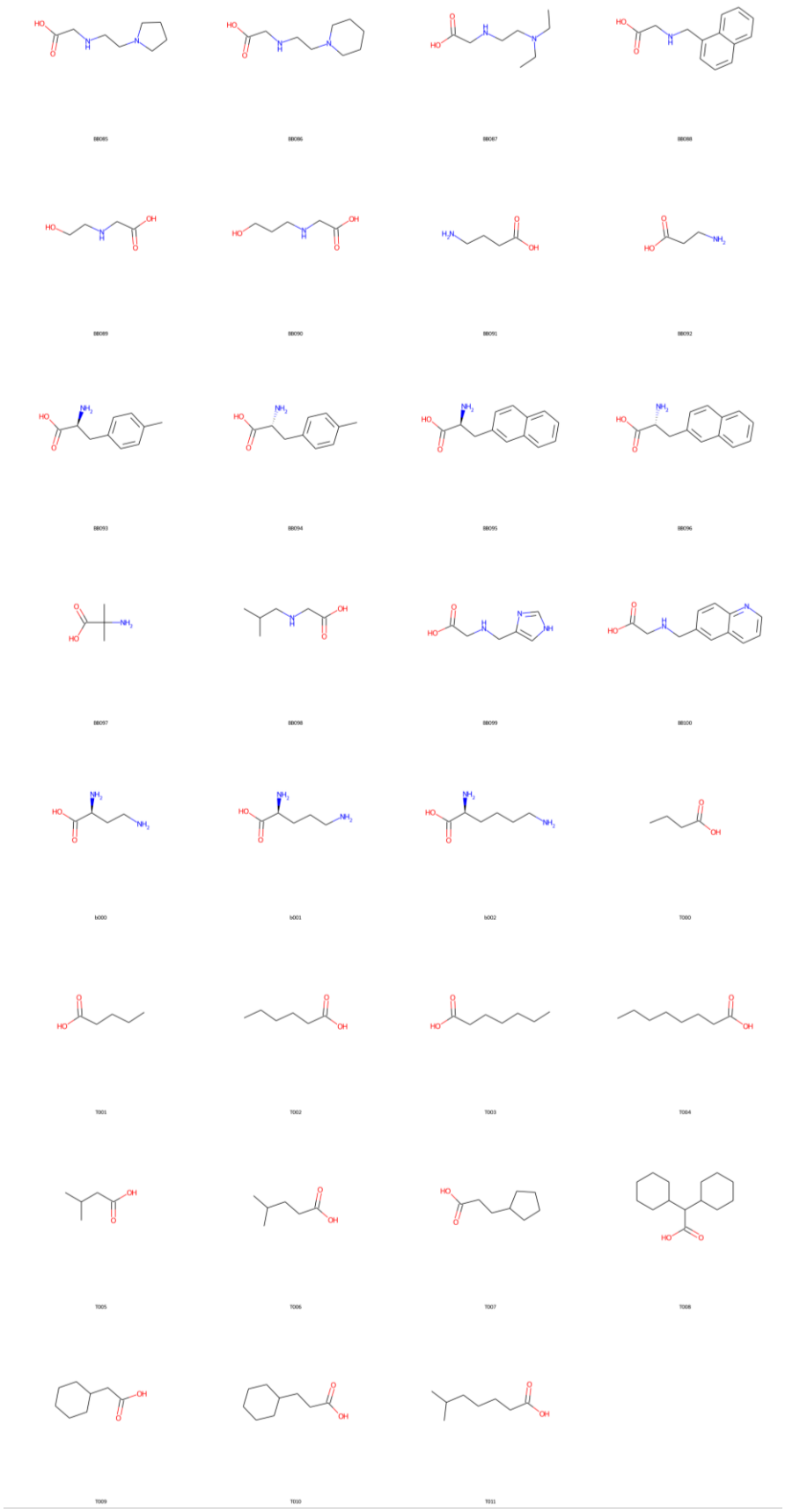


Figure C1. Structures of the building blocks used by the PDGA.

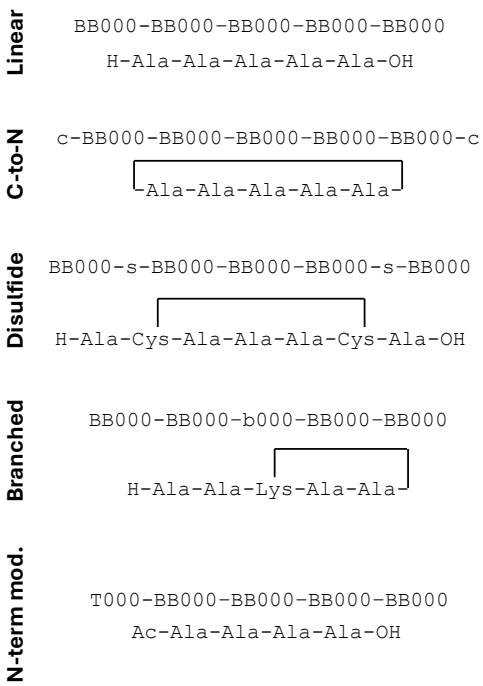


Figure C2. Linear format used to store sequences in the PDGA with examples for all possible types of modifications. Modifications that are compatible with each other can also be combined, e.g. C-to-N and disulfide cyclization.

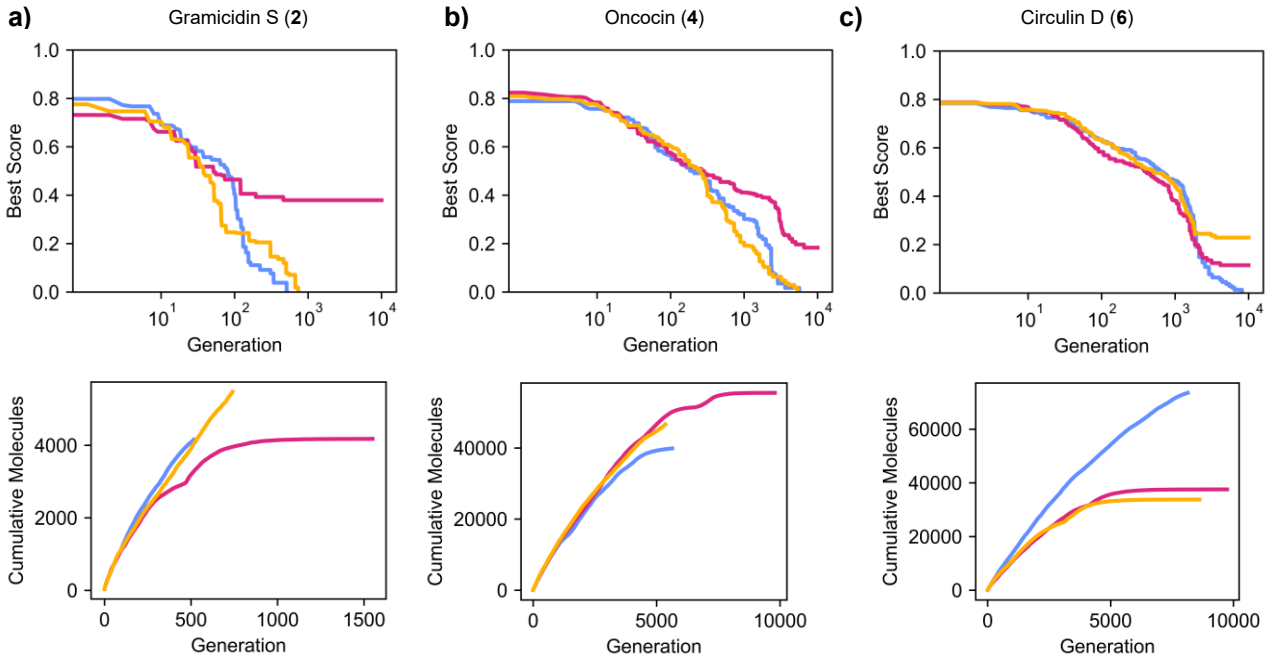


Figure C3. Analysis of three parallel PDGA runs starting from 50 random sequences towards selected queries. Top plots show the overall best score throughout the trajectory; the bottom plots show the cumulative number of unique new molecules generated throughout the trajectory for a) gramicidin S, b) oncocin, and c) circulin D.

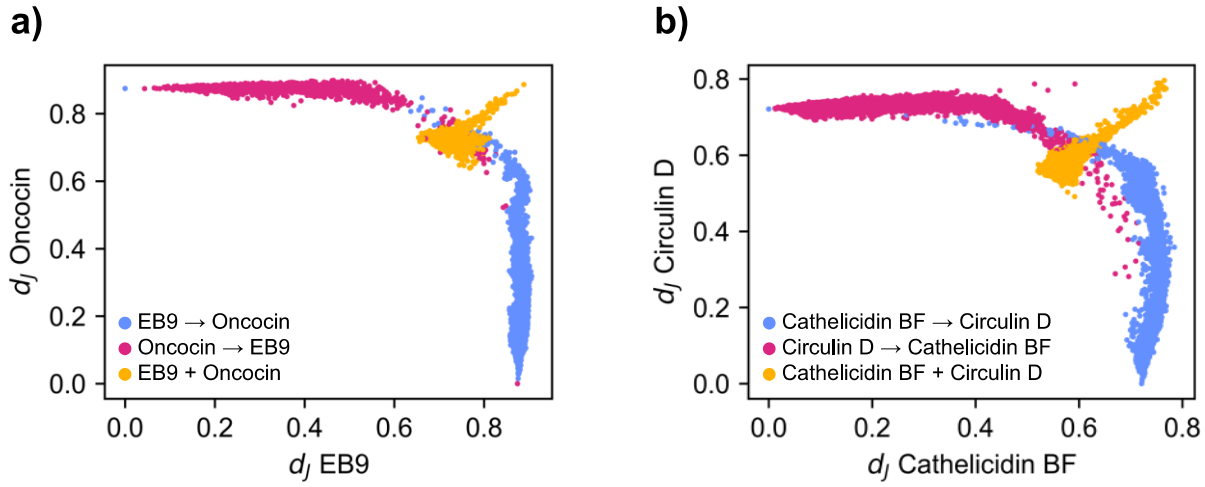
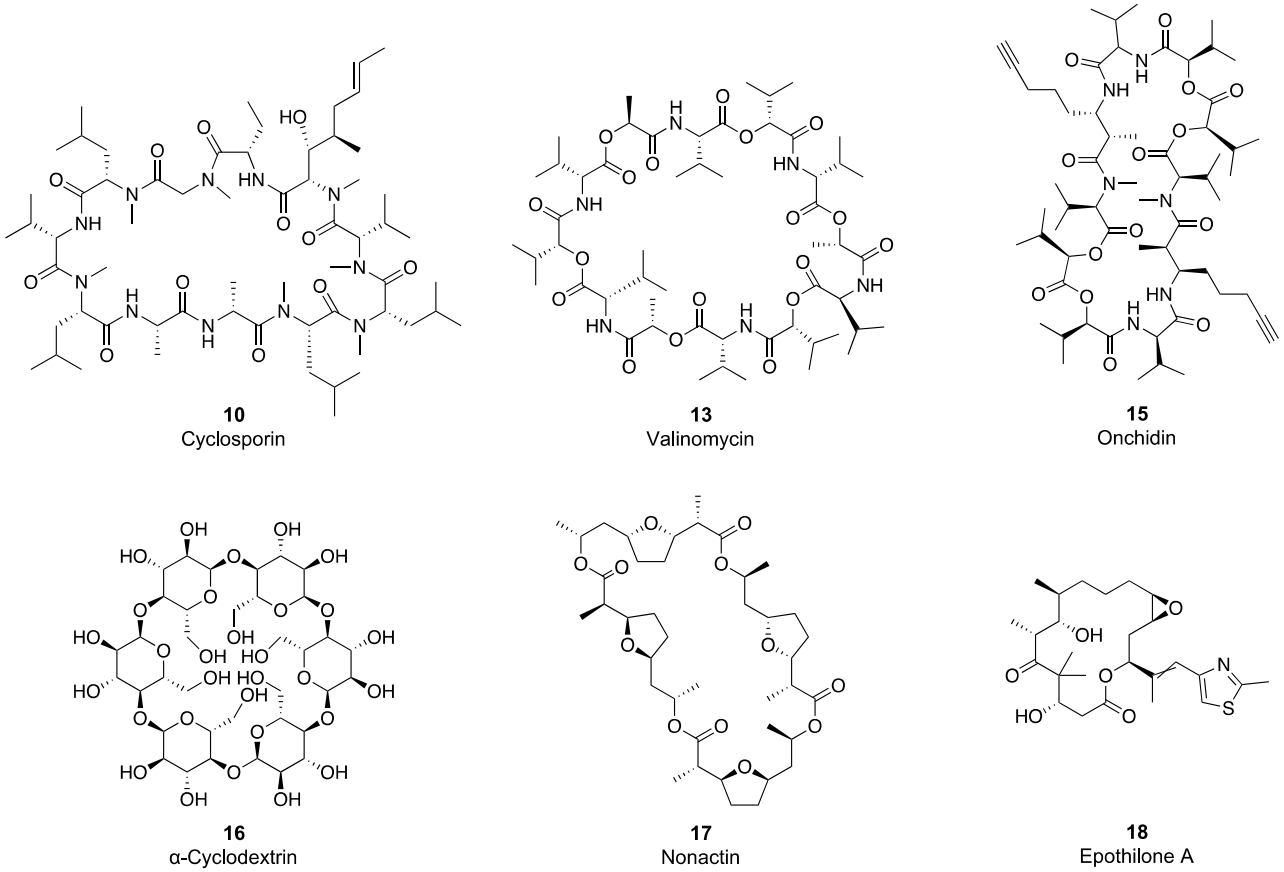


Figure C4. Jaccard distance of molecules selected from the different traversal trajectories towards a) oncocin and EB9 and b) circulin D and cathelicidin BF.



10
Cyclosporin

13
Valinomycin

15
Onchidin

16
α-Cyclodextrin

17
Nonactin

18
Epothilone A

Figure C5. Structures of the non-peptide macrocycle queries for the PDGA runs using the MXFP similarity as fitness function.

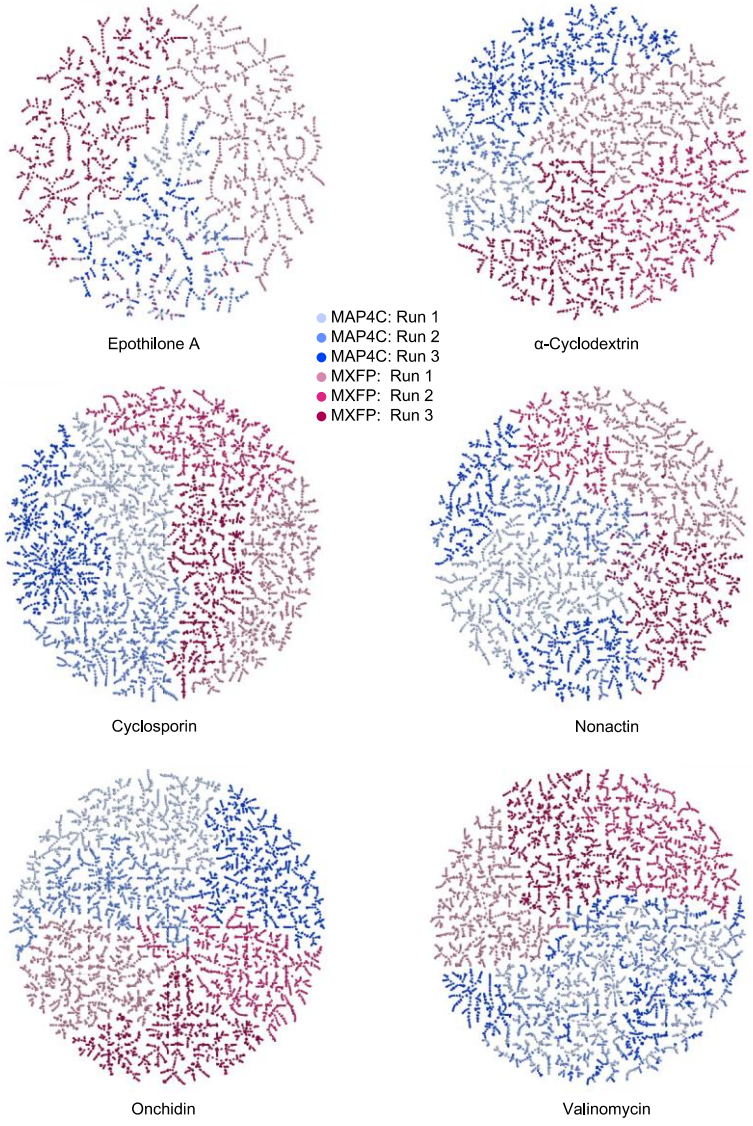


Figure C6. TMAPs of top 1000 molecules generated in each of three parallel MAP4C and MXFP trajectories of selected non-peptide queries. Interactive TMAPs: https://tm.gdb.tools/map4/10E60/epothilone_tmap.html

https://tm.gdb.tools/map4/10E60/cyclodextrin_tmap.html

https://tm.gdb.tools/map4/10E60/cyclosporin_tmap.html

https://tm.gdb.tools/map4/10E60/nonactin_tmap.html

https://tm.gdb.tools/map4/10E60/onchidin_tmap.html

https://tm.gdb.tools/map4/10E60/valinomycin_tmap.html

Appendix D - Supporting information for: Can large language models predict antimicrobial peptide activity and toxicity?

Table D1. Performance metrics of all models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold for activity and hemolysis separately. Results for reduced training sets are reported for 20% and 2% size of the original activity dataset and 10% of the original hemolysis set.

Model	ROC AUC	Accuracy	Precision	Recall	F1
GPT-3 Ada act.	0.84	0.78	0.78	0.78	0.78
GPT-3 Babbage act.	0.85	**0.79**	**0.79**	0.78	**0.79**
GPT-3 Curie act.	**0.86**	**0.79**	0.78	**0.81**	**0.79**
GPT-3 Ada 20% act.	0.75	0.69	0.7	0.67	0.68
GPT-3 Babbage 20% act.	0.76	0.69	0.7	0.69	0.68
GPT-3 Curie 20% act.	0.76	0.7	0.71	0.71	0.71
GPT-3 Ada 2% act.	0.66	0.6	0.6	0.63	0.61
GPT-3 Babbage 2% act.	0.66	0.62	0.6	0.73	0.66
GPT-3 Curie 2% act.	0.65	0.6	0.6	0.63	0.61
GPT-3 Ada hem.	**0.9**	0.82	0.8	**0.79**	0.79
GPT-3 Babbage hem.	0.87	0.8	0.76	0.76	0.76
GPT-3 Curie hem.	0.89	**0.84**	**0.82**	**0.79**	**0.8**
GPT-3 Ada 10% hem.	0.72	0.68	0.63	0.58	0.6
GPT-3 Babbage 10% hem.	0.72	0.7	0.65	0.6	0.62
GPT-3 Curie 10% hem.	0.73	0.68	0.63	0.59	0.61

Table D2. Mean and standard deviation of performance metrics of selected models tested on antimicrobial activity and hemolysis classification. The best value for each metric is highlighted in bold.

Model	ROC AUC	Accuracy	Precision	Recall	F1
Features SVM act.	0.65 ± 0.01	0.65 ± 0.01	0.65 ± 0.01	0.63 ± 0.01	0.64 ± 0.01
MAP4C SVM act.	0.8 ± 0.01	**0.8 ± 0.01**	**0.78 ± 0.01**	0.83 ± 0.01	**0.80 ± 0.01**
RNN act.	**0.85 ± 0.01**	0.78 ± 0.01	0.76 ± 0.02	0.81 ± 0.01	0.78 ± 0.01
GPT-3.5 Turbo act.	0.69 ± 0.01	0.69 ± 0.01	0.62 ± 0.01	**0.95 ± 0.01**	0.75 ± 0.01
Features SVM hem.	0.62 ± 0.01	0.64 ± 0.01	0.59 ± 0.02	0.48 ± 0.02	0.53 ± 0.01
MAP4C SVM hem.	0.82 ± 0.02	**0.82 ± 0.01**	**0.78 ± 0.02**	**0.82 ± 0.04**	**0.79 ± 0.01**
RNN hem.	**0.87 ± 0.01**	0.81 ± 0.01	0.77 ± 0.03	0.79 ± 0.03	0.78 ± 0.01
GPT-3.5 Turbo hem.	0.47 ± 0.01	0.48 ± 0.01	0.38 ± 0.02	0.36 ± 0.02	0.37 ± 0.02

Table D3. Training times and costs of GPT models on the full training sets.

Model	Time (h)	Costs ($)
GPT-3 Ada Activity	01:05:04	$0.39
GPT-3 Babbage Activity	01:09:38	$0.59
GPT-3 Curie Activity	01:15:05	$2.93
GPT-3.5 Turbo Activity	00:53:24	$7.00
GPT-3 Ada Hemolysis	00:55:37	$0.09
GPT-3 Babbage Hemolysis	00:57:19	$0.13
GPT-3 Curie Hemolysis	01:08:09	$0.67
GPT-3.5 Turbo Hemolysis	00:55:58	$1.66

Appendix E - Supporting information for: Assigning the stereochemistry of natural products by machine learning

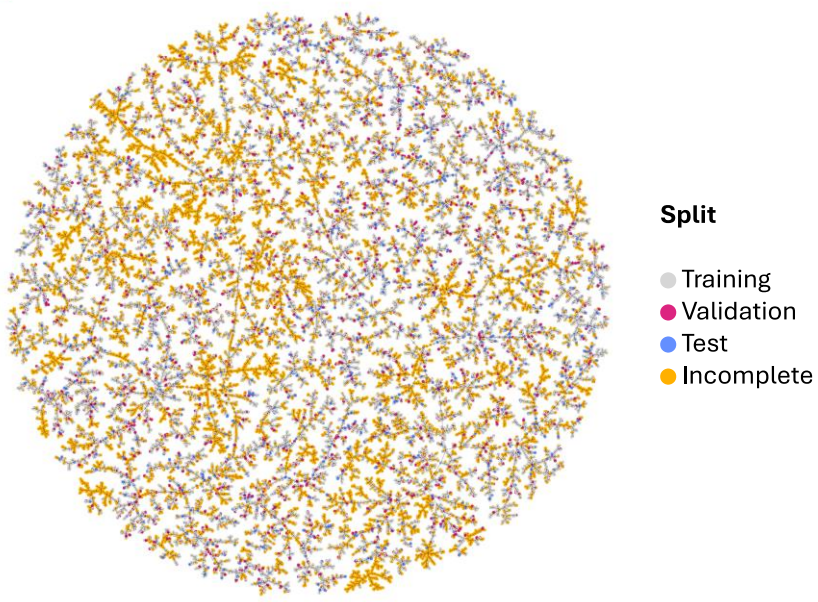


Figure E1. MAP4C TMAP of 103,605 unique compounds with an associated DOI, extracted from the COCONUT database. The TMAP visualization is colored according to the data split (grey-training, red-validation, blue-test) for SMILES with complete stereocenter assignment (63,988) and as incomplete (yellow) for SMILES with icomplete stereocenter assignment (39,617).

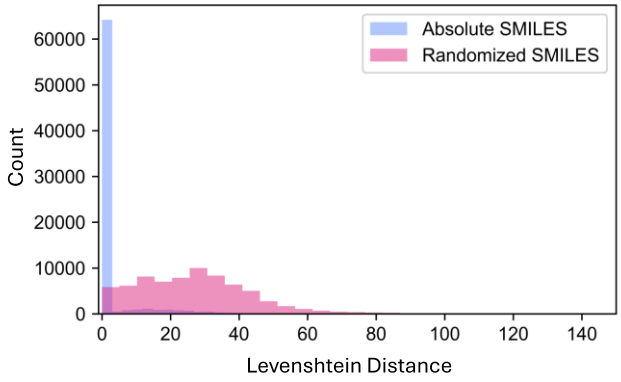


Figure E2. Levenshtein distance distribution comparing absolute canonical SMILES generated with RDKit to two alternatives: (1) absolute canonical SMILES with stereochemistry-defining characters removed (blue) and (2) absolute randomized SMILES generated with RDKit (red).

Declaration of consent

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

Name/First Name: Orsi Markus

Registration Number: 16-946-220

Study program: Chemistry and Molecular Sciences

Bachelor ☐ Master ☐ Dissertation ☒

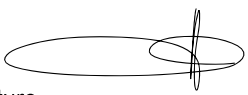
Title of the thesis: Computational Strategies for the Data-Driven Discovery of Antimicrobial Peptides

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, 01.04.2025

Place/Date



Signature