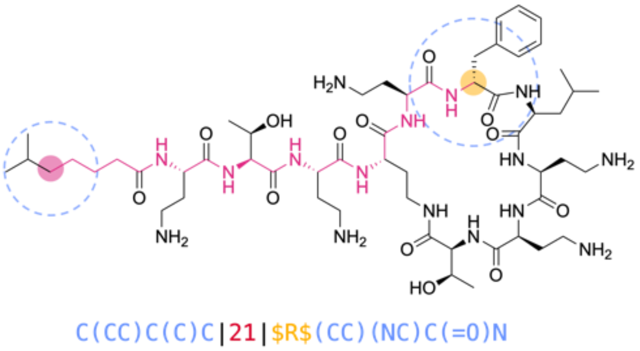
# One chiral fingerprint to find them all

Markus Orsi,[a] and Jean-Louis Reymond[a]*

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

*e-mail:* *jean-louis.reymond@unibe.ch*



## Abstract

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

**Results:** 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 as good as the achiral MAP4, ECFP and AP fingerprints in non-stereoselective virtual screening benchmarks. Furthermore, it readily distinguishes between thousands of stereoisomers in complex natural products and peptides.

**Conclusion:** MAP4C is recommended as a generally applicable chiral molecular fingerprint.

**Keywords**: molecular fingerprints, stereochemistry, virtual screening, chemical space, atom-pairs

## Introduction

Many computational tasks related to small molecule drug discovery, such as similarity searches,[1,2] target prediction,[3–7] ligand-based virtual screening[8] and visualization of large databases of drug-like molecules,[9–18] can be performed using vectors encoding molecular structure, called molecular fingerprints.[19,20] 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,[21–23] 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,[24,25] in our case for peptides with variable chain topology and stereochemistry,[26–28] we have adapted molecular fingerprints based on atom-pairs[29–32] for large molecules such as peptides and proteins.[33–35] In particular, we combined atom-pair analysis and circular substructures as encoded the Morgan fingerprint ECFP4,[36,37] with the principle of data compression using MinHashing,[38–41] 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.[42] These pairs are written in the form of two canonicalized SMILES[43,44] 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)[37] and Atom Pair (AP)[29] fingerprints fall short. In addition, MAP4 outperforms these and many other fingerprints in virtual screening benchmarks for both small molecule drugs[20] and peptides.[42]

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,[45] COCONUT,[46] and ZINC,[47] are chiral (**Figure 1a**). 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 1b/c**), as well as double bond stereochemistry.

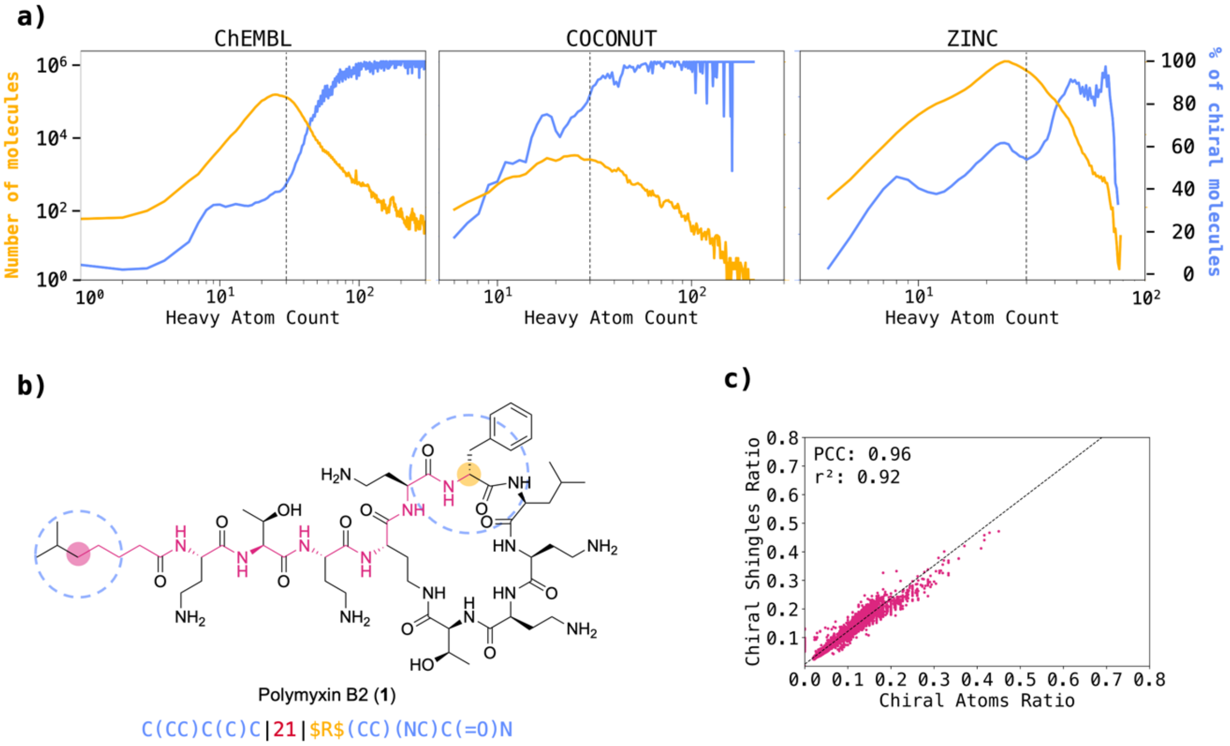


**Figure 1:** Molecular chirality and fingerprints. (**a**) Correlation between chirality and heavy atom count (HAC) across ChEMBL, COCONUT, and ZINC datasets. The blue line depicts the percentage of chiral molecules relative to HAC. A steady increase in the percentage of chiral molecules is observed 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.

# Methods

### 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.[38,42]

### Benchmark

The virtual screening performance of the MAPC fingerprint was evaluated in a comparative study with commonly used fingerprints (ECFP4,[37] ECFP6,[37] Atom-Pair[29]) in a benchmark adapted from Riniker and Landrum.[20] Since the structure SMILES in the original benchmark do not contain any stereochemistry, the respective chiral SMILES (when applicable) were retrieved from the DUD,[48] MUV[49] and ChEMBL[45] 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.[42]

For every set, 5 randomly selected actives were extracted and stored in a separate file. 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.

### Stereoisomers, isomers and scrambled sequences

We enumerated all possible stereoisomers of molecules **1 – 14** (Figure 1c and Figure 4) 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.[50,51] 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, yelding 746 unique structures for **15** and 126 for **16**. For all stereoisomers and permutations, fingerprints were calculated as 2048-bit vectors.

### 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 neighbors in the MAP4C feature space were extracted from the LSH forest and used to calculate the TMAP layout.[16] The resulting layout was displayed in an interactive TMAP using the open-source Faerun package.[15]

## Results and Discussion

### 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.[38–41] 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.[42] 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 1b**).

When applied to a dataset of chiral molecules uniformly sampled from the Riniker and Landrum benchmark (**Figure S1**),[20] 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 S2a**), MAP4C (largest diameter of four bonds, **Figure 1c**) and MAP6C (largest diameter of six bonds, **Figure S2b**). 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.

## 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-neighbor 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 physico-chemical 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,[20] which considers 118 active and decoy datasets taken from DUD,[48] MUV,[49] and ChEMBL.[45] 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),[42] for which we challenge the fingerprint to retrieve BLAST search analogs.[52]

Both of these benchmarks tested the ability of the fingerprints to retrieve bioactive analogs without consideration of stereochemistry. 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). 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 non-stereoselective benchmarks (**Figure 2a/b** and **Figure S3-S7**). 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 2c**).

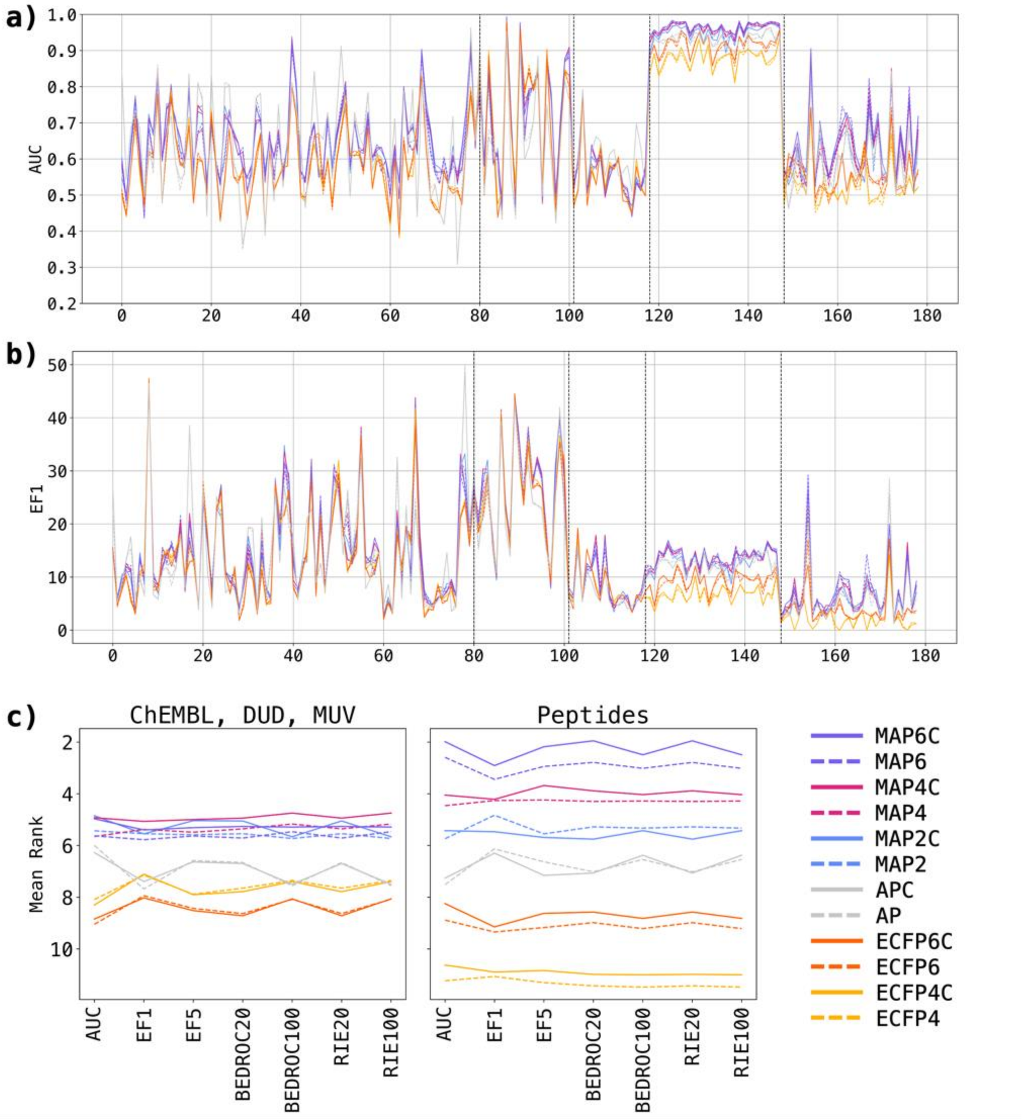


**Figure 2**: Virtual Screening benchmark. **a**) AUC and **b**) EF1 of MAP6 (purple), MAP4 (magenta), MAP2 (blue), AP (grey), ECFP6 (orange) and ECFP4 (yellow) and across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled 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.

## 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 comprizing carbohydrates, peptides and macrocyclic natural products containing up to thousands of stereoisomers per molecule (**Figure 3** and **Table 1**).

For carbohydrates, the six tested chiral fingerprints 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**), and all but APC distinguished the 16,384 stereoisomers of the aminoglycoside antibiotic validamycin A (**5**). However, MAP2C, MAP4C and MAP6C were the only fingerprints capable of differentiating the nine possible stereoisomers of the signaling carbocyclic sugar *myo*-inositol (**6**), which all gave identical ECFP4C, ECFP6C and APC fingerprints.

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,[28] 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**),[53,54] as well as the 4096 stereoisomers of polymyxin B2 (**1**), used as last resort antibiotic against multidrug resistant bacteria.[55] 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**).[56] 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.

| **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.

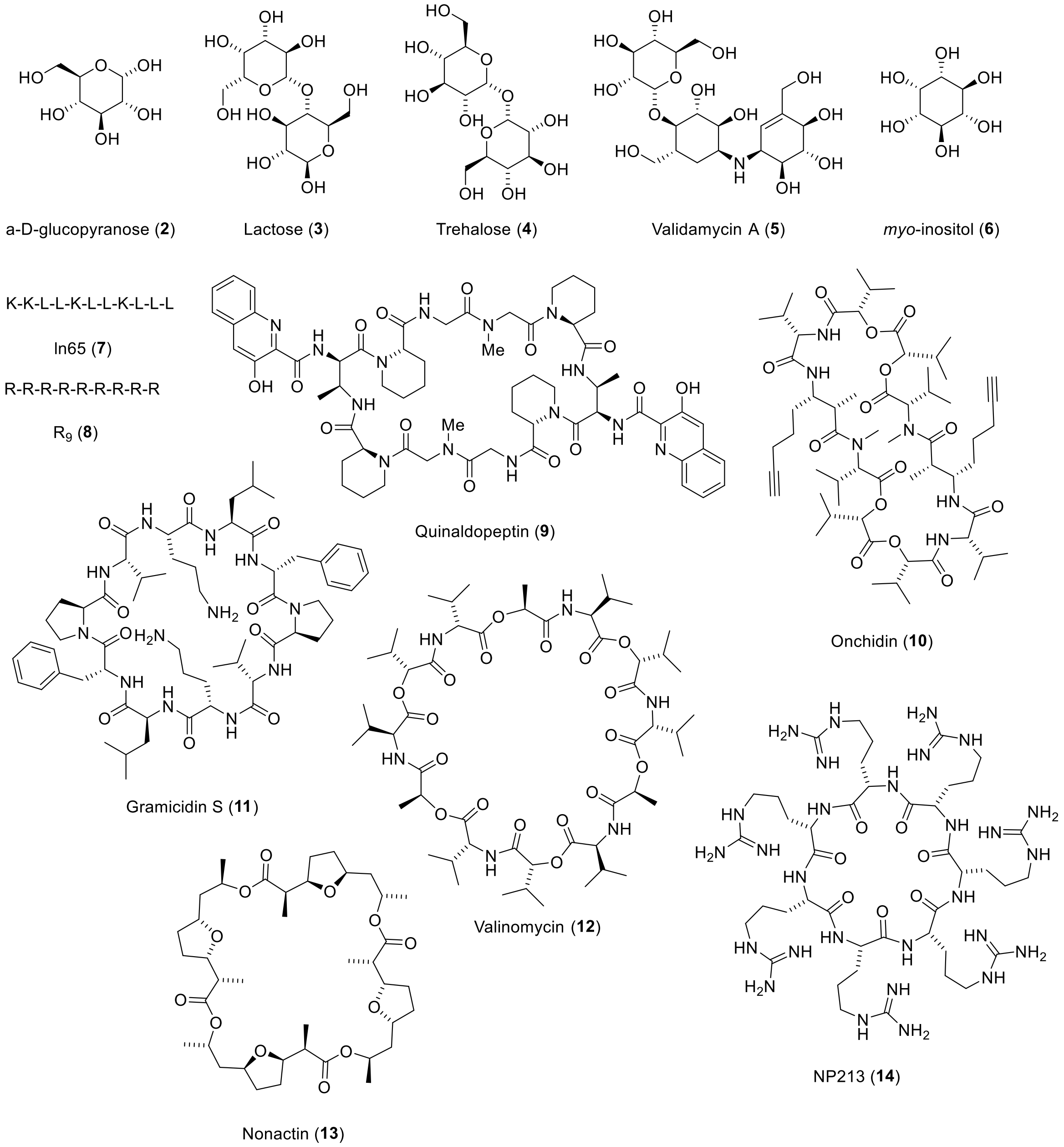


**Figure 3.** Structures of natural products selected for the stereoisomer distinction task.

## 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,[57,58] and between all pairs of stereoisomers in the set. We similarly analyzed all pairs involving the 48 structural isomers of 4-aminopiperazine (**16**), a similar drug scaffold, and the stereoisomeric pairs within the set. In both cases, all six fingerprints ranked pairs stereoisomers closer to each other than pairs of structural isomers (**Figure 4a/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 4c/d** and **Figure S8/9**). 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 5a/b**).[16] 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.

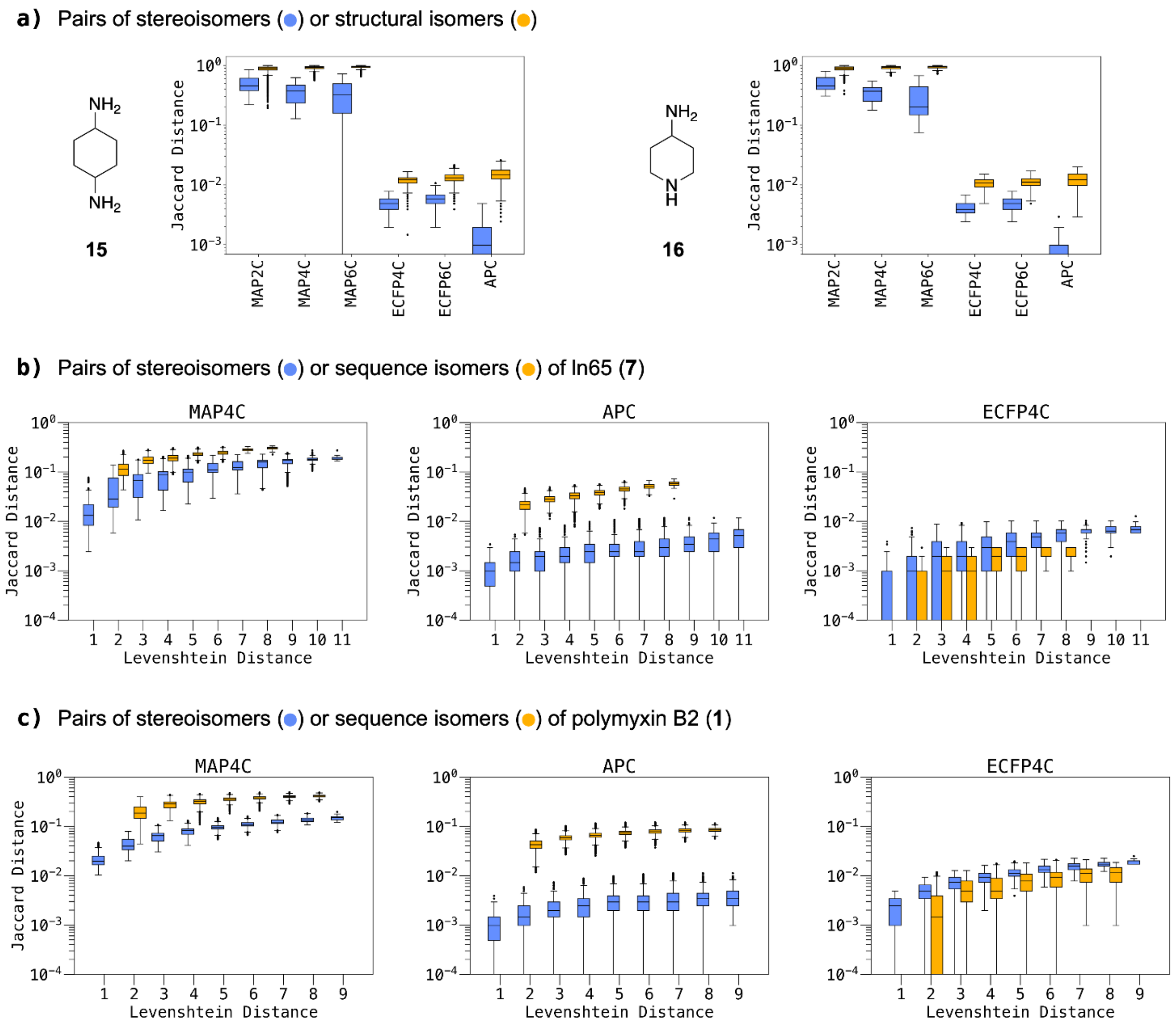


**Figure 4**: Differentiation between stereoisomers and structural isomers, shown as box plots of average Jaccard distances between pairs of stereoisomers (blue) or structural/sequence isomers (yellow). **a**) structural isomers of 1,4-diaminocyclohexane (203) and 4-aminopiperidine (48) and their diastereomers **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 Figures S10 and S11 for plots with MAP6C, MAP2C and ECFP6C. See methods for details.

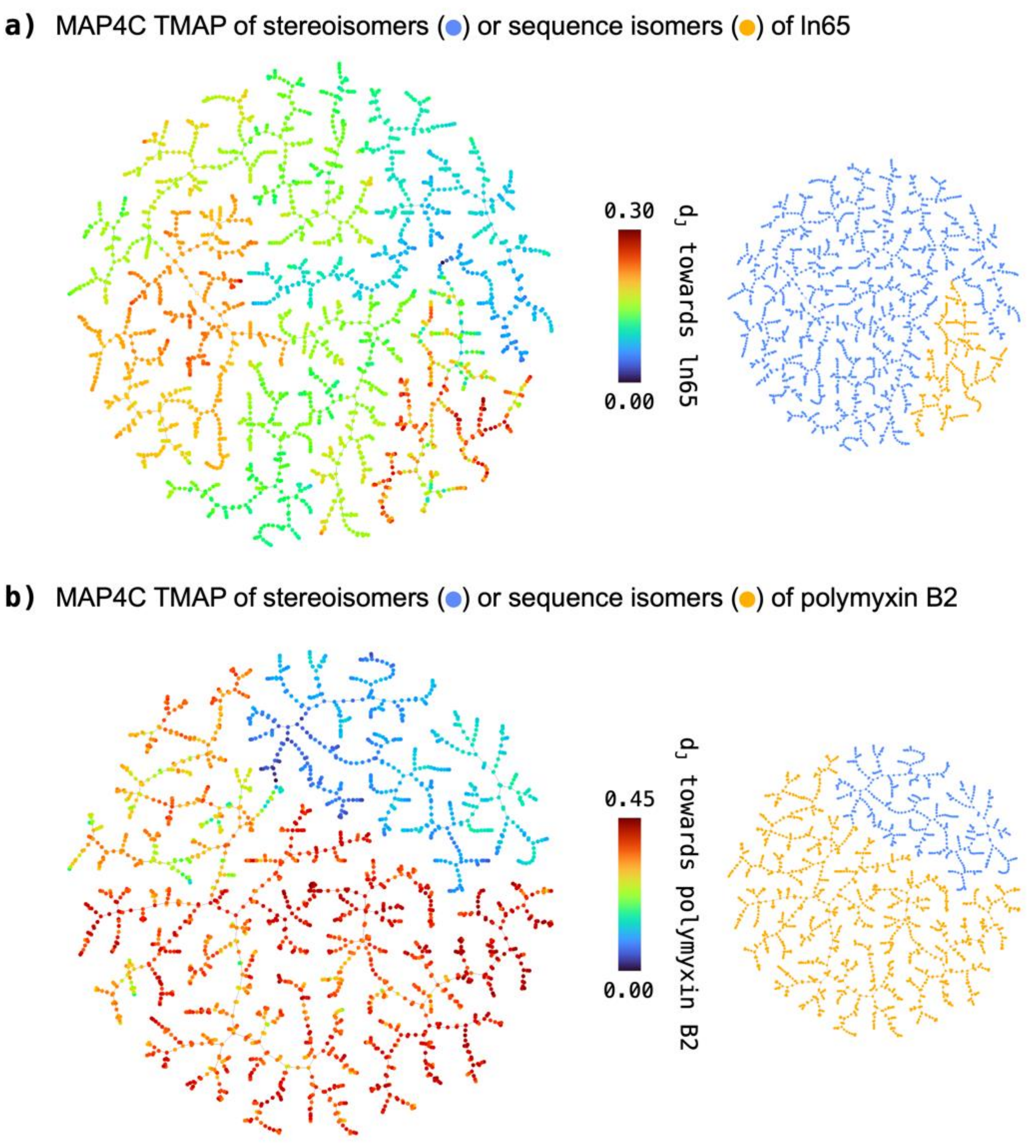


**Figure 5**: MAP4C TMAPs showing the Jaccard distance ($d_J$; rainbow) of stereoisomers (blue) and sequence isomers (yellow) towards their respective queries: (a) ln65, 2,048 diastereomers and 330 sequence isomers. The interactive version of the TMAP is accessible under https://tm.gdb.tools/map4/MAP4C_ln65/ (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/.

## Conclusions

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.

## List of abbreviations

**AP(C):** Atom pair fingerprint (chiral)
**AUC:** Area under the curve
**BEDROC:** Boltzmann-enhanced discrimination of the receiver operating characteristic
**BLAST:** Basic local alignment search tool
**CIP:** Cahn-Ingold-Prelog
**DUD:** Directory of useful decoys
**ECFP(C):** Extended connectivity fingerprint (chiral)
**EF:** Enrichment factor
**FP**: Fingerprint
**GDB:** Generated database
**HAC:** Heavy atom count
**JD:** Jaccard distance
**LSH:** Locality sensitive hashing
**MAP(C):** MinHash atom pair fingerprint (chiral)
**MQN:** Molecular quantum numbers
**MUV:** Maximum unbiased validation data sets
**PMB2:** Polymyxin B2
**RIE:** Robust initial enhancement
**Ro5:** Rule of fives
**ROC:** Receiver operating characteristic
**SMILES:** Simplified molecular-input line-entry system
**TMAP:** Tree map

# Declarations

### 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.

### Competing interests

The authors declare that they have no competing interests.

### Authors' 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.

### Acknowledgements

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

## References

(1) 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.

(2) 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.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) Blackmond, D. G. The Origin of Biological Homochirality. *Cold Spring Harb Perspect Biol* **2019**, *11* (3), a032540. https://doi.org/10.1101/cshperspect.a032540.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

(30) 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.

(31) 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*, 1892–1897. https://doi.org/10.1021/ci500232g.

(32) Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminf.* **2015**, *7*, 3.

(33) 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.

(34) Capecchi, A.; Awale, M.; Probst, D.; Reymond, J. L. PubChem and ChEMBL beyond Lipinski. *Mol. Inf.* **2019**, *38*, 1900016. https://doi.org/10.1002/minf.201900016.

(35) 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.

(36) 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.

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

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

(39) 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.

(40) Manber, U. Finding Similar Files in a Large File System. In *Usenix Winter 1994 Technical Conference*; 1994; pp 1–10.

(41) 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.

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

(43) 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.

(44) 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.

(45) 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.

(46) 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.

(47) 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.

(48) 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.

(49) 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.

(50) 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.

(51) 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.

(52) 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.

(53) 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.

(54) 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.

(55) 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.

(56) 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.

(57) 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.

(58) 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.

# Supplementary Information for:

# One chiral fingerprint to find them all

Markus Orsi[a] and Jean-Louis Reymond[a]*

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

*e-mail: jean-louis.reymond@unibe.ch*

## Table of Contents

1. Supplementary figures

Figure S1 ........................................................................................2

Figure S2 ........................................................................................2

Figure S3 ........................................................................................3

Figure S4 ........................................................................................3

Figure S5 ........................................................................................4

Figure S6 ........................................................................................4

Figure S7 ........................................................................................5

Figure S8 ........................................................................................6

Figure S9 ........................................................................................7

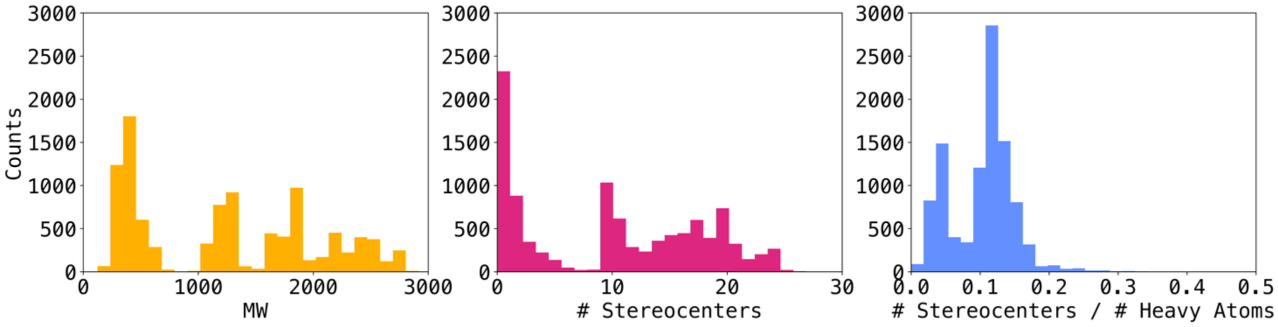


**Figure S1**. Distribution of molecular weight (MW) (yellow), number of stereocenters (magenta) and ratio of stereocenters to heavy atom count (blue) in the set uniformly sampled from the extended benchmark. The set contained a total of 10,122 compounds and was used to determine the relative impact of stereochemistry encoding on total similarity.

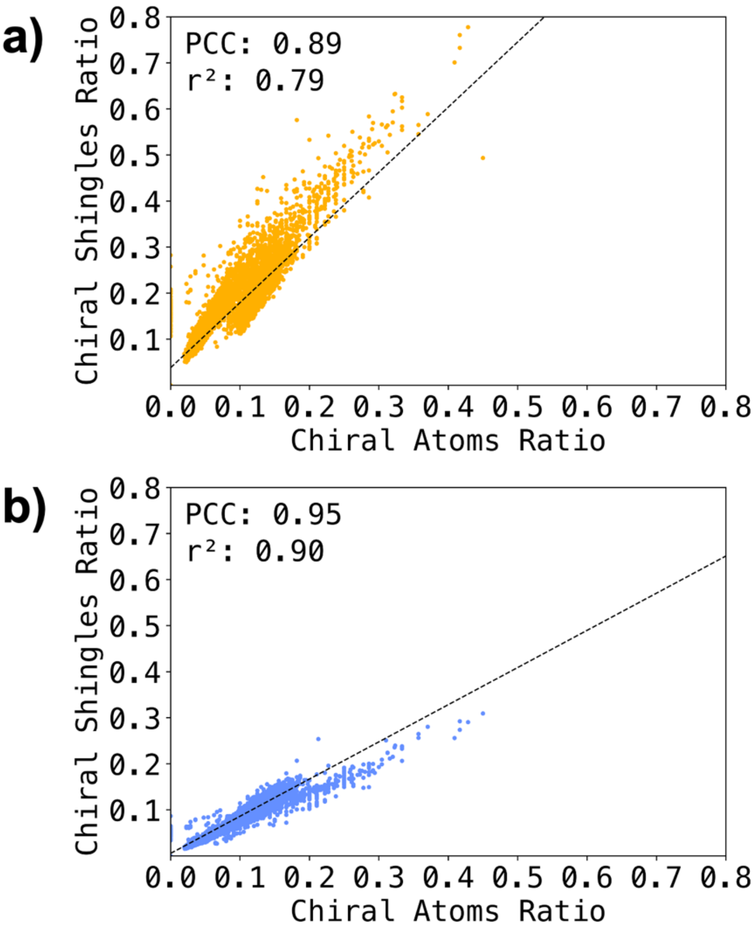


**Figure S2**. Scatterplots of chiral shingle ratio vs. chiral atoms ratio for a) radius = 1 b) radius = 2 and c) radius = 3. Additionally, the $r^2$ of the linear fit and the Pearson correlation coefficient (PCC) are reported. All reported PCCs are statistically significant.

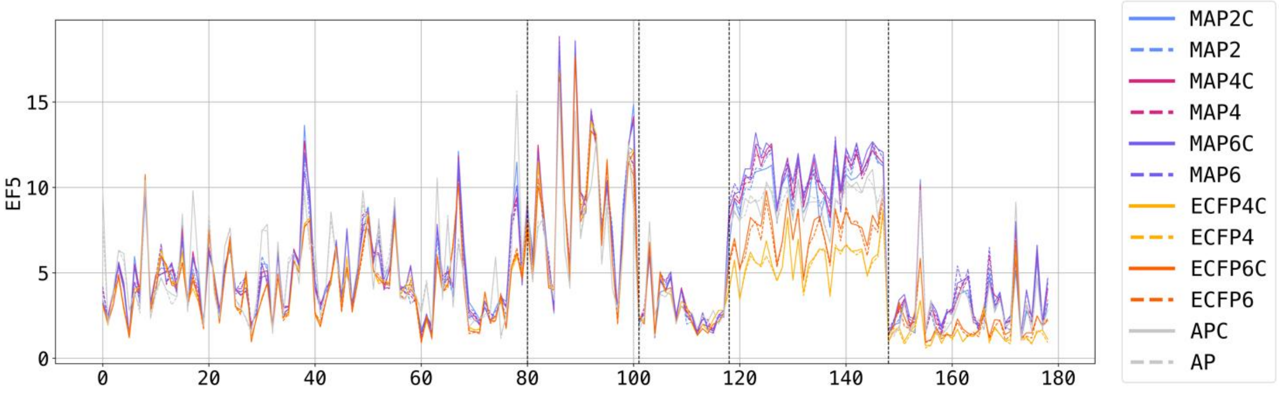


**Figure S3**. EF5 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

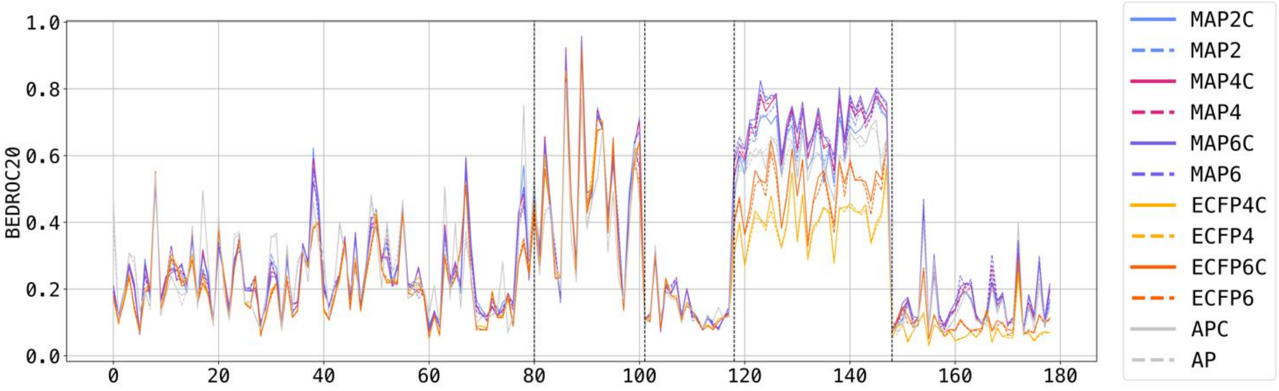


**Figure S4**. BEDROC20 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are 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.

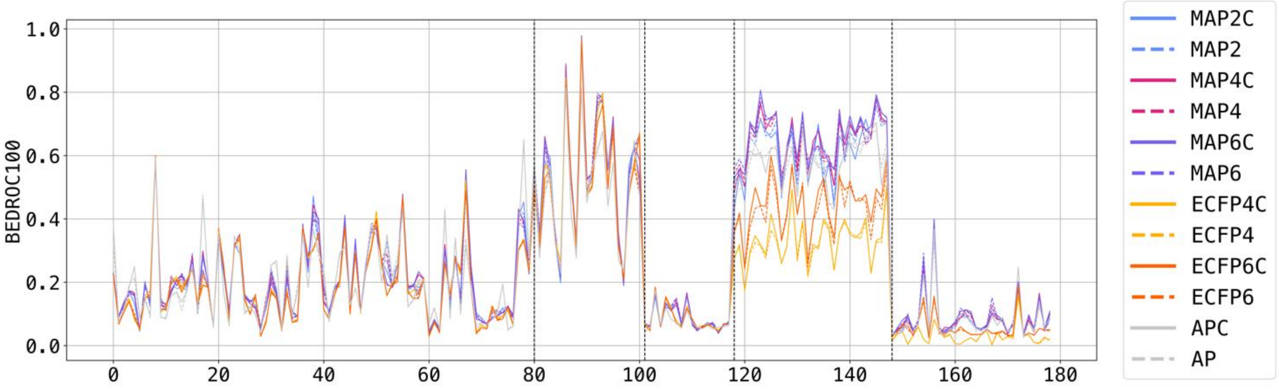


**Figure S5**. BEDROC100 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are 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.

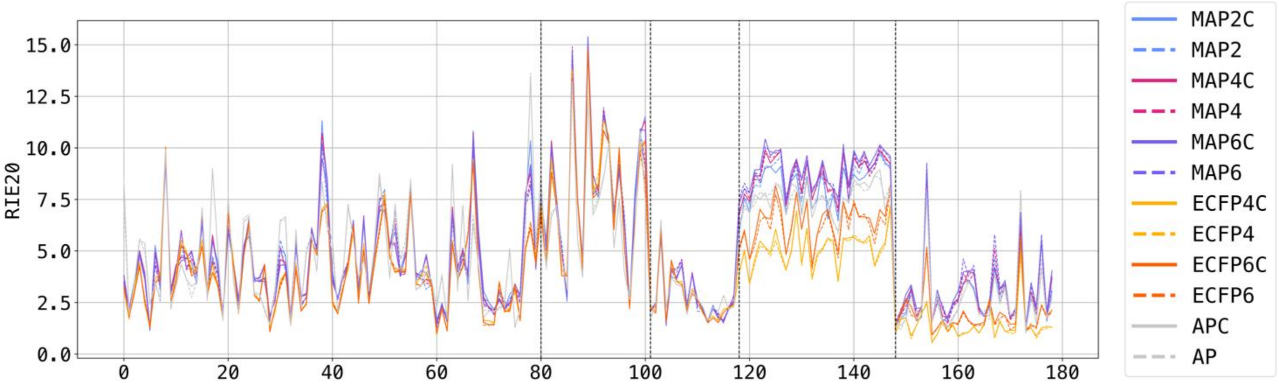


**Figure S6**. RIE20 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

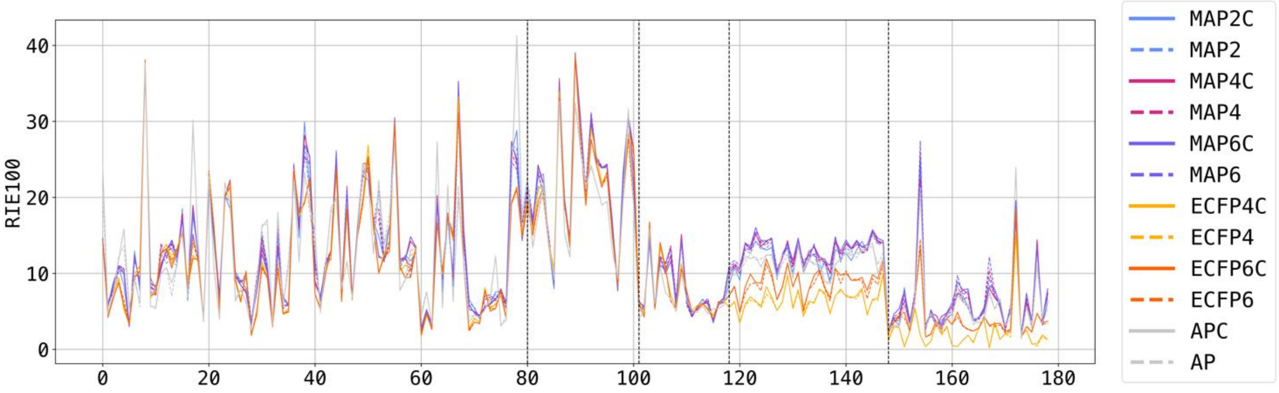


**Figure S7**. RIE100 of MAP2 (blue), MAP4 (magenta), MAP6 (purple), AP (grey), ECFP4 (yellow) and ECFP6 (orange) across all small molecules and peptide targets (80 ChEMBL targets, 21 DUD targets, 17 MUV targets, 30 mutated peptide targets, and 30 scrambled peptide targets). Chiral fingerprints are displayed as bold lines, non-chiral fingerprints are displayed as dashed lines. The value displayed for each dataset is the mean metric of 5 runs.

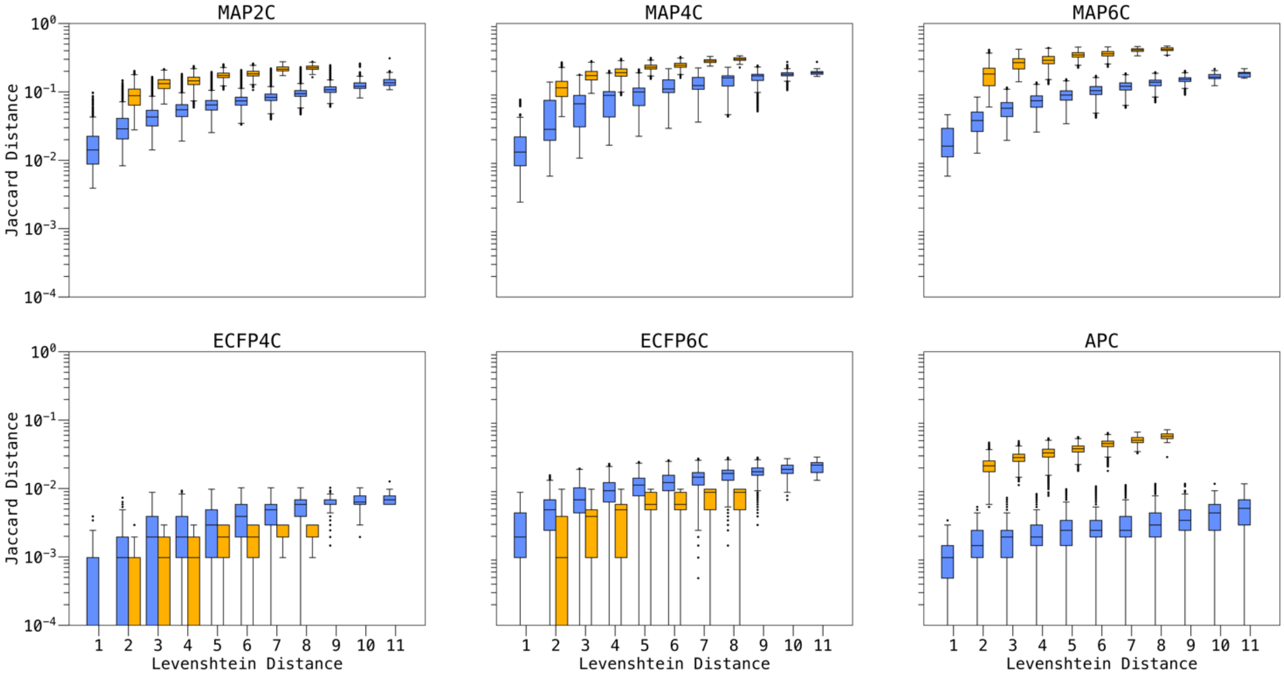


**Figure S8**. Comparative analysis of MAP2C, MAP4C, MAP6C, APC, ECFP4C and ECFP6C Jaccard distance assignment on ln65 diastereomers (blue) and structural isomers (yellow). The distance distributions are grouped by Levenshtein distance, used to determine the number of mutations from any sequence to ln65. 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.

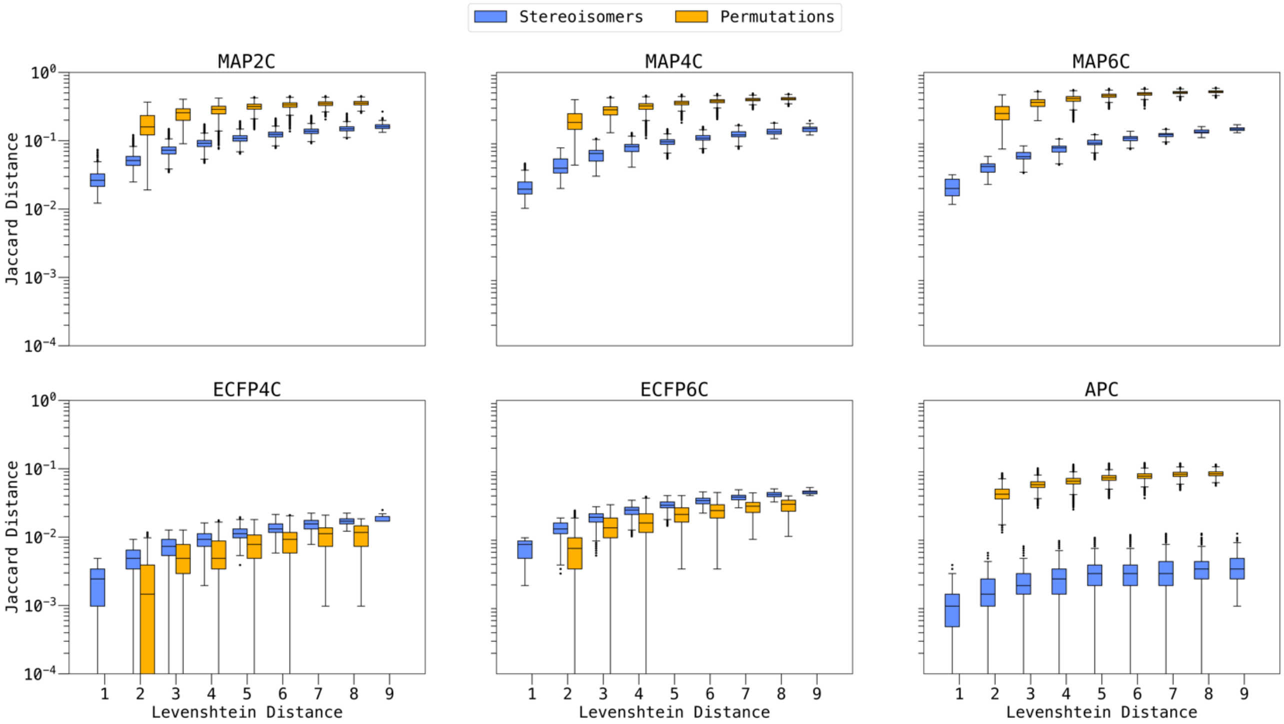


**Figure S9**. Comparative analysis of MAP2C, MAP4C, MAP6C, APC, ECFP4C and ECFP6C Jaccard distance assignment on polymyxin B2 diastereomers (blue) and structural isomers (yellow). The distance distributions are grouped by Levenshtein distance, used to determine the number of mutations from any sequence 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.