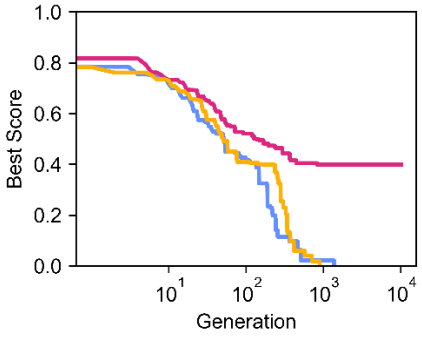
Navigating a 1E+60 Chemical Space

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

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 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 molecules otherwise difficult to design. We also show that PDGA can be used to generate median molecules and analogs of non-peptide target molecules.

Keywords: therapeutic peptides, chemical space, cheminformatics, genetic algorithm

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,[1–3] drug discovery has been fascinated and partly driven by large numbers.[4–6] Approaches ranged from the "needle in a haystack" method of high-throughput screening typical for genetically encoded display libraries[7,8] and DNA-encoded libraries,[9,10] to the concept of chemical space guiding the design of focused libraries of small drug-like molecules,[11–13] fragments[14,15] and peptides.[16–18] 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.[19–22] Despite of being rather constrained, this oligomer chemical space has proven amenable to virtual screening and sufficiently diverse to solve most drug discovery problems,[23–25] probably because biomolecules are themselves oligomers and their binding sites are usually suitable for partly flexible, pearl-string like molecules.[26–28]

Following up on our interest for exhaustive enumeration of chemical space,[29–31] 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.[6,32] We also aimed to demonstrate virtual screening at that library size focusing on ligand-based virtual screening (LBVS).[33,34] LBVS consists in identifying analogs of a reference bioactive compound by scoring the virtual library using molecular similarity measures such as molecular fingerprints,[35–37] or shape-based comparisons.[38–42] As discussed below, we achieved our goals for the case of mixed peptide-peptoids accessible by solid-phase peptide synthesis (SPPS),[43] 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),[44] which evolves analogs of any target molecule by searching a topologically diverse oligomer space using molecular fingerprint similarity as fitness function, and

can be used to design new analogs of known peptides as recently demonstrated experimentally for antimicrobial peptide dendrimers.[45] Specifically, we computed the fitness function using the macromolecule extended atom pair fingerprint (MXFP)[46,47] and the chiral MinHashed atom pair fingerprint (MAP4C),[48,49] both designed for large molecules.

Methods

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)[50] 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 S1**).[51,52] 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. Both, the enhanced sequence format, and the corresponding SMILES, are stored in the results files.

Genetic Algorithm

We modified our previously reported PDGA[44] by computing fitness functions either as the Jaccard distance (d_J) to the target molecule computed using the molecular fingerprint MAP4C,[49] 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,[47] here saving only molecule

with $d_{CBD} \leq 300$ as trajectory molecules. 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 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.

Results and Discussion

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 1,** 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)[53] or the diabetes/obesity drug semaglutide (31 residues).[54] 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),[50] which can be easily assembled by SPPS with the sub-monomer approach.[55] All 100 building blocks selected were commercially available or easily accessible in a protected from suitable for peptide and/or peptoid submonomer SPPS (**Figure S1**).

Table 1. Influence of oligomer length M and number of building blocks N on virtual library size.

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**

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[56] and which favor cellular uptake in natural products[57] and extend peptide circulation times via albumin binding.[58] We also added several options for cyclization (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 1**, 3rd column).[59]

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 1**).[44] 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.

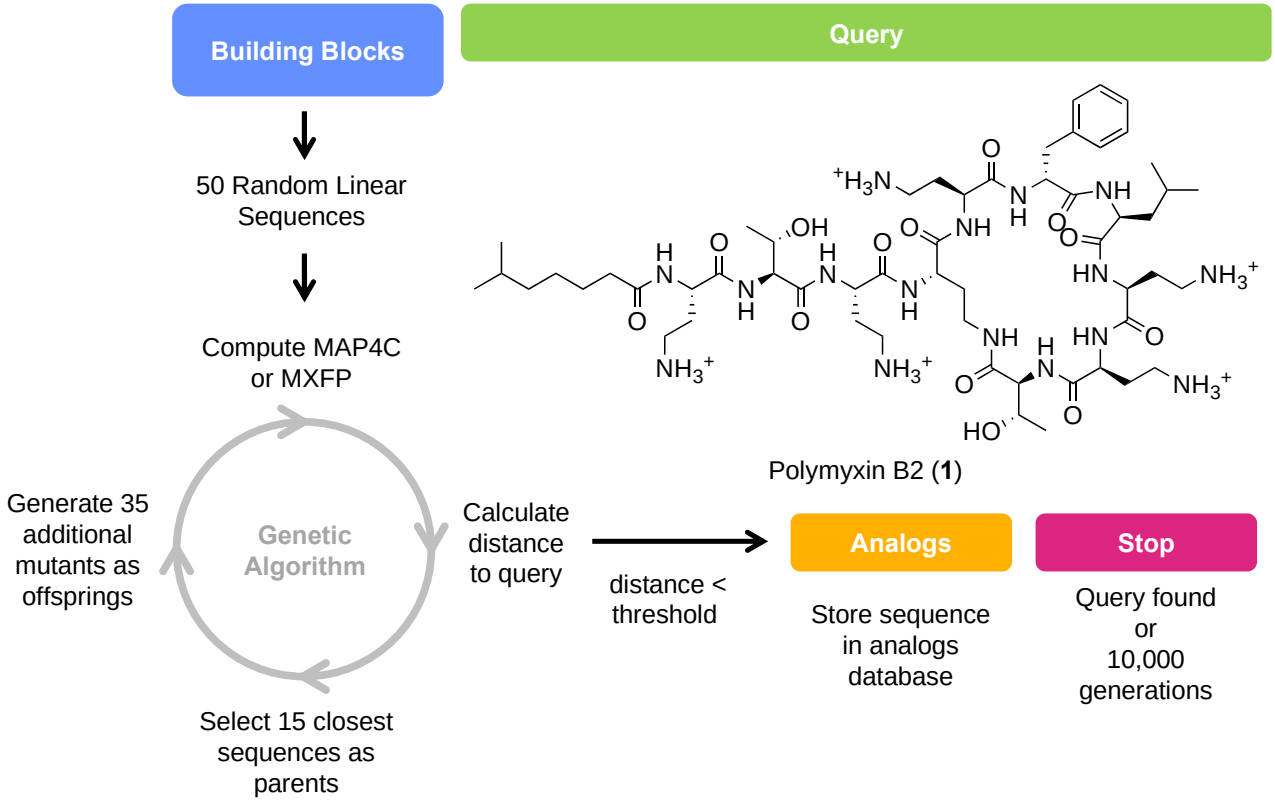


Figure 1. Design of PDGA. PDGA uses a list of input building blocks to generate a set of random linear sequences. The sequences are encoded using either the MAP4C or MXFP fingerprints. The fingerprints are used to determine the fitness of the sequences by calculating the distance towards a specified 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),[56] gramicidin S (**2**, 10 residues, antimicrobial),[60,61] the mixed peptide/peptoid hybrid EB9 (**3**, 11 residues, antibacterial),[43] oncocin (**4**, 19 residues, antimicrobial),[62] cathelicidin BF (**5**, 30 residues,

immunomodulatory peptide),[63] and circulin D (**6**, 30 residues, anti-HIV)[64](**Figure S2**). 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.[48,49]

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 2**). 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 2** and **S3**, 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 2** and **S3**, 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.

Table 2: Results of three parallel PDGA runs for queries **1-6**.

Query	length	Structure[a)]	# generations to query[b)]			# unique structures (% with $d_J < 0.5$)			# unique structures not counting diastereomers (% with $d_J < 0.5$)		
			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[c)]	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) PDGA was run on the linear sequence lacking the cystine bridges.

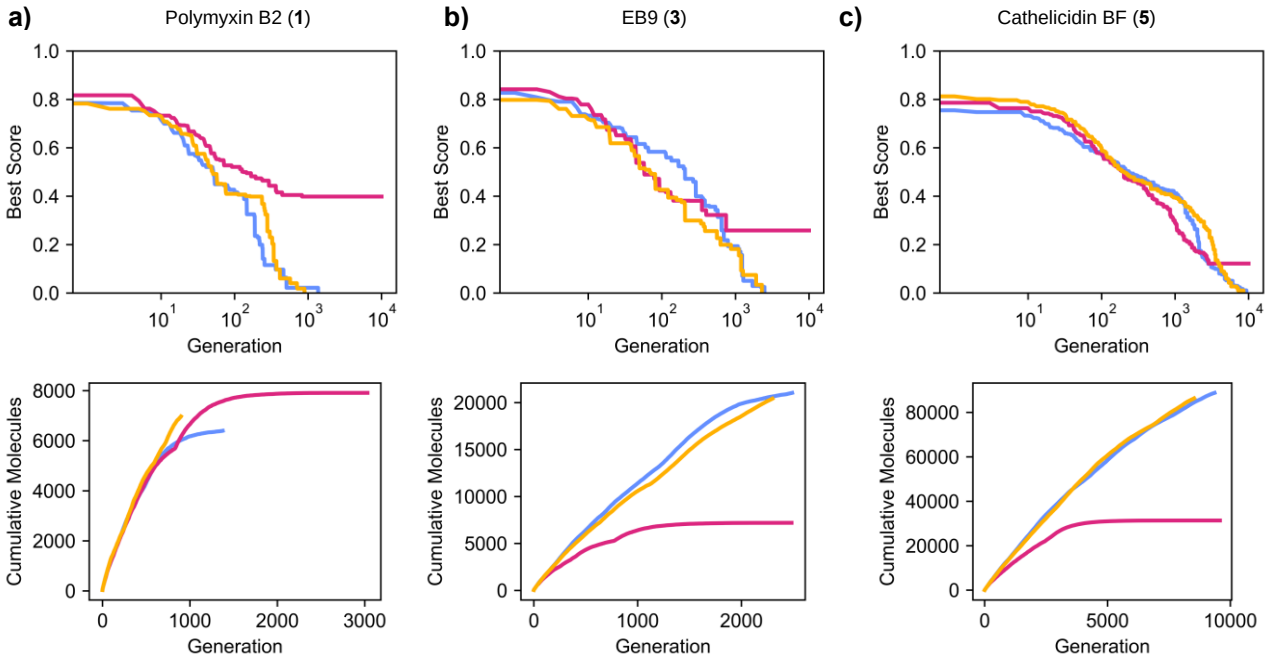


Figure 2. Analysis of three parallel PDGA runs starting from 50 random sequences towards selected queries. Top plots show the overall best score throughout the trajectory; the bottom plots show the cumulative number of unique new molecules generated throughout the trajectory for **a)** polymyxin B2, **b)** EB9, and **c)** cathelicidin BF.

To get a closer insight into the analogs generated by PDGA, we focused on the case of polymyxin B2 (**Figure 3**). 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 3a**). 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 3b**). Note that 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 3c**).

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 3d**). 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)[65] 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 3d**).

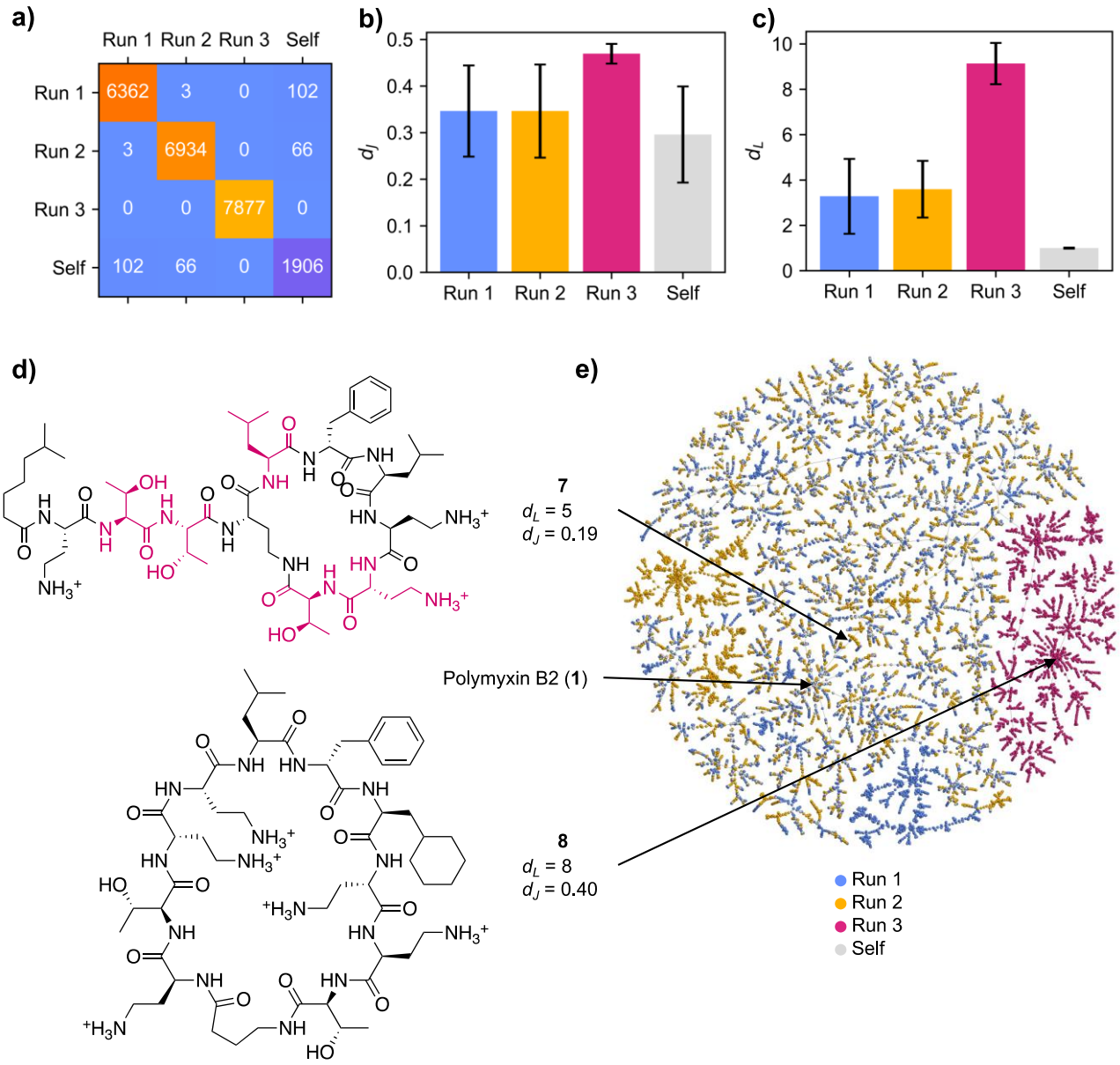


Figure 3. Analysis of polymyxin B2 runs starting from 50 random linear sequences (Run 1-3) or from polymyxin B2 without stopping condition (Self). **a)** Heatmap indicating the number of generated compounds with $d_J < 0.5$ to polymyxin B2 for each trajectory, along with the number of overlapping compounds. **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.

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,[66,67] 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 4a** and **S4**).

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 4a** and **S4**). 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 4b/c**).

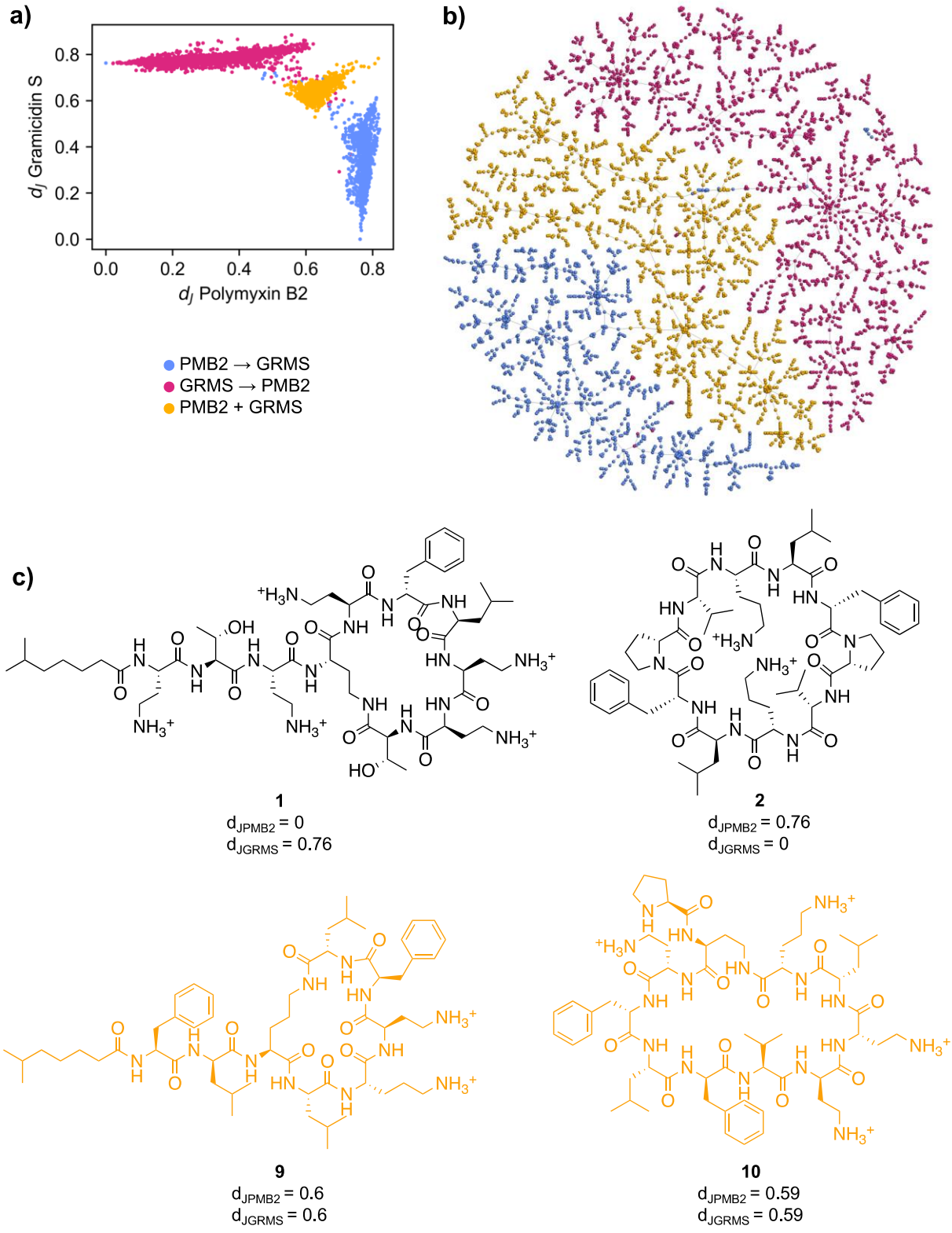


Figure 4. Visualization of traversal trajectories and median molecules between polymyxin B2 and gramicidin S. **a)** Jaccard distance of molecules selected from the different trajectories towards polymyxin B2 and gramicidin S. The trajectory from polymyxin B2 to gramicidin S is displayed in blue, the 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.

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 S5**). 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 5** and **S6**). For instance, in the case of cyclosporin (**11**), which contains several N-methylated amide bonds essential for its membrane permeability, 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.

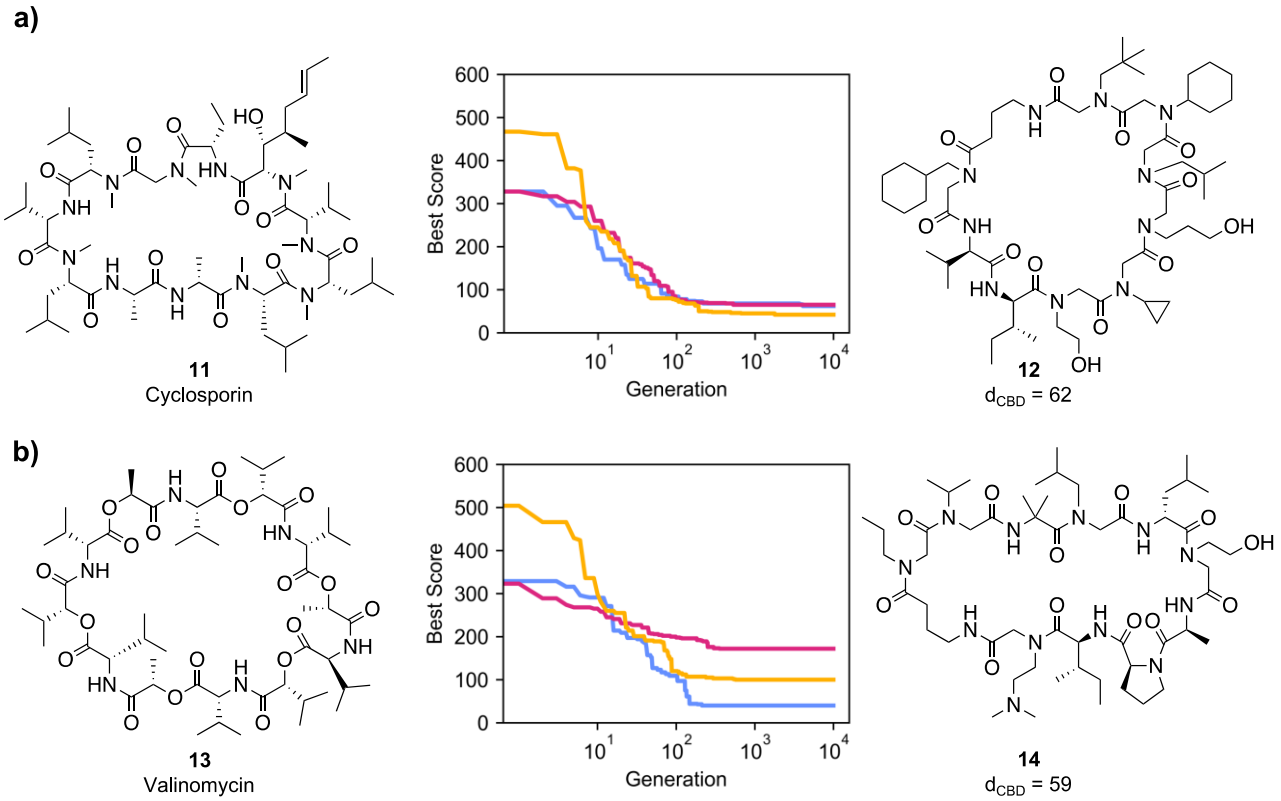


Figure 5. Non-peptide macrocycles, the overall best score throughout the trajectories and the corresponding best scoring MXFP analog from three combined runs for **a)** cyclosporin and **b)** valinomycin. The MXFP d_{CBD} is reported for each analog. See also Figure S6 for further details.

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 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.

Although our PDGA sometimes failed to converge on a target molecule by getting stuck in local minima, the computational expense to correct this problem by introducing a duplicate molecule check at every iteration is far too large, and one is much better served by running the algorithm several

times. 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. Ongoing studies along these lines will be reported separately.

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.

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.

Acknowledgements

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

References

(1) 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.

(2) 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.

(3) 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.

(4) 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.

(5) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat Rev Drug Discov.* **2003**, *2* (5), 369–378.

(6) Kirkpatrick, P.; Ellis, C. Chemical Space. *Nature* **2004**, *432* (7019), 823–823.

(7) 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.

(8) 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.

(9) 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.

(10) Dockerill, M.; Winssinger, N. DNA-Encoded Libraries: Towards Harnessing Their Full Power with Darwinian Evolution. *Angew. Chem.* **2023**, *135* (9), e202215542. https://doi.org/10.1002/ange.202215542.

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

(12) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432* (7019), 855–861. https://doi.org/10.1038/nature03193.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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. *ChemComm* **2018**, *54*, 5130–5133. https://doi.org/10.1039/c8cc02412j.

(18) 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.
(19) 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. *J Chem Inf Model* **2015**, *55* (9), 1824–1835. https://doi.org/10.1021/acs.jcim.5b00203.
(20) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today* **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.
(21) 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.
(22) 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.
(23) Irwin, J. J.; Gaskins, G.; Sterling, T.; Mysinger, M. M.; Keiser, M. J. Predicted Biological Activity of Purchasable Chemical Space. *J Chem Inf Model* **2018**, *58* (1), 148–164. https://doi.org/10.1021/acs.jcim.7b00316.
(24) 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.
(25) 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.
(26) Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7* (9), 1884–1897. https://doi.org/10.1002/pro.5560070905.
(27) Kandel, J.; Tayara, H.; Chong, K. T. PUResNet: Prediction of Protein-Ligand Binding Sites Using Deep Residual Neural Network. *J. Cheminformatics* **2021**, *13* (1), 65. https://doi.org/10.1186/s13321-021-00547-7.
(28) 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.
(29) 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.
(30) 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.
(31) 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.
(32) 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.
(33) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-Based Virtual Screening. *Drug Discov. Today* **2011**, *16* (9), 372–376. https://doi.org/10.1016/j.drudis.2011.02.011.
(34) 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.

(35) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11* (23–24), 1046–1053.
(36) 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.
(37) 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.
(38) 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.
(39) 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.
(40) 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.
(41) 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.
(42) Awale, M.; Jin, X.; Reymond, J. L. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J Cheminf* **2015**, *7*, 3.
(43) 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.
(44) 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.
(45) 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.
(46) 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.
(47) Orsi, M.; Probst, D.; Schwaller, P.; Reymond, J.-L. Alchemical Analysis of FDA Approved Drugs. *Digit. Discov.* **2023**, *2* (5), 1289–1296. https://doi.org/10.1039/D3DD00039G.
(48) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.
(49) Orsi, M.; Reymond, J.-L. One Chiral Fingerprint to Find Them All. ChemRxiv December 20, 2023. https://doi.org/10.26434/chemrxiv-2023-33j02.
(50) Zuckermann, R. N. Peptoid Origins. *Pept. Sci.* **2011**, *96* (5), 545–555. https://doi.org/10.1002/bip.21573.
(51) 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.
(52) 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.
(53) 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.

(54) 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.
(55) 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.
(56) 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.
(57) 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.
(58) 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. *J Pharm Sci* **1997**, *86* (12), 1365–1368. https://doi.org/10.1021/js9701768.
(59) 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.
(60) 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.
(61) 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. *Int. J. Pept. Protein Res.* **1996**, *47* (6), 460–466. https://doi.org/10.1111/j.1399-3011.1996.tb01096.x.
(62) 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.
(63) 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. Pharm.* **2015**, *12* (5), 1648–1661. https://doi.org/10.1021/acs.molpharmaceut.5b00069.
(64) 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.
(65) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminformatics* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.
(66) 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.
(67) van Deursen, R.; Reymond, J.-L. Chemical Space Travel. *ChemMedChem* **2007**, *2* (5), 636–640. https://doi.org/10.1002/cmdc.200700021.

Supplementary Information for:

Navigating a 1E+60 Chemical Space

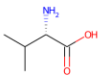
Markus Orsi[a] and Jean-Louis Reymond[a]*

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

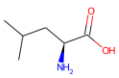
e-mail: jean-louis.reymond@unibe.ch

Table of Contents

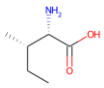
BB001



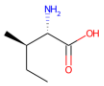
BB002



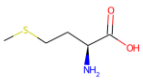
BB003



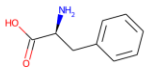
BB004



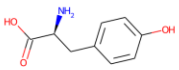
BB005



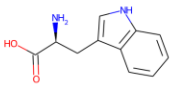
BB006



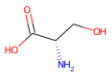
BB007



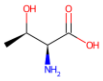
BB008



BB009

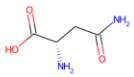


BB010

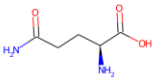


BB011

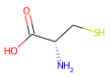
BB012



BB013



BB014

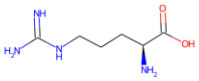


BB015

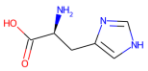
BB016

BB017

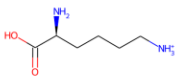
BB018



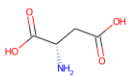
BB019



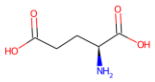
BB020



BB021

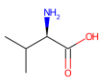


BB022

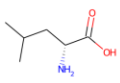


BB023

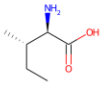
BB024



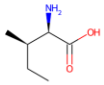
BB025



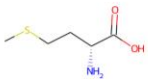
BB026



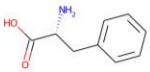
BB027



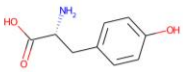
BB028



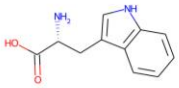
BB029



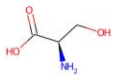
BB030



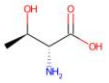
BB031



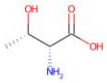
BB032



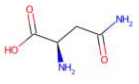
BB033



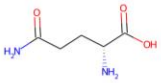
BB034



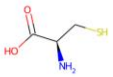
BB035



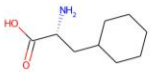
BB036



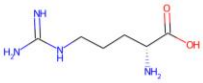
BB037



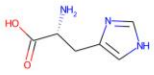
BB038



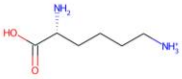
BB039



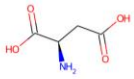
BB040



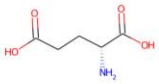
BB041



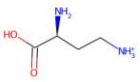
BB042



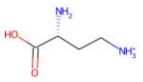
BB043



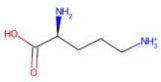
BB044



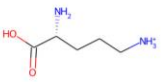
BB045



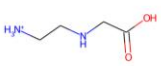
BB046



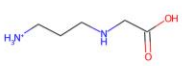
BB047



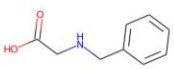
BB048



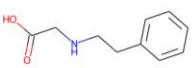
BB049



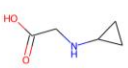
BB050



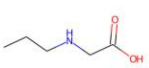
BB051



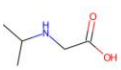
BB052



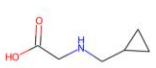
BB053



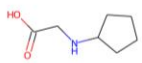
BB054



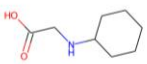
BB055



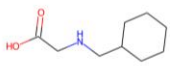
BB056



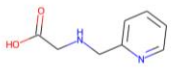
BB057



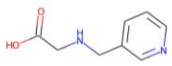
BB058



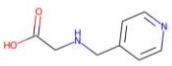
BB059



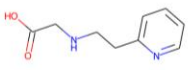
BB060



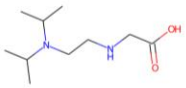
BB061



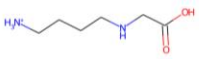
BB062



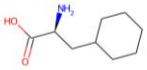
BB063



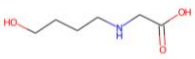
BB064



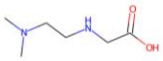
BB065



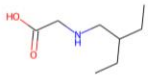
BB066



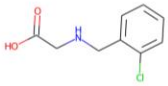
BB067



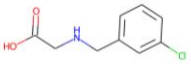
BB068



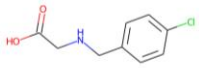
BB069



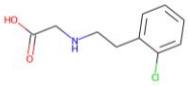
BB070



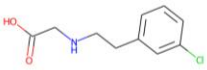
BB071



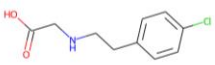
BB072



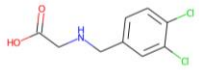
BB073



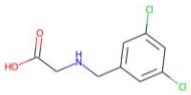
BB074



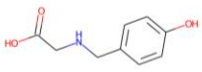
BB075



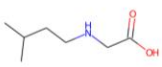
BB076



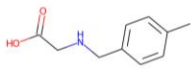
BB077



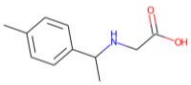
BB078



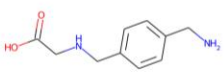
BB079



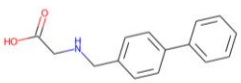
BB080



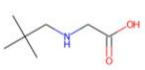
BB081



BB082



BB083



BB084

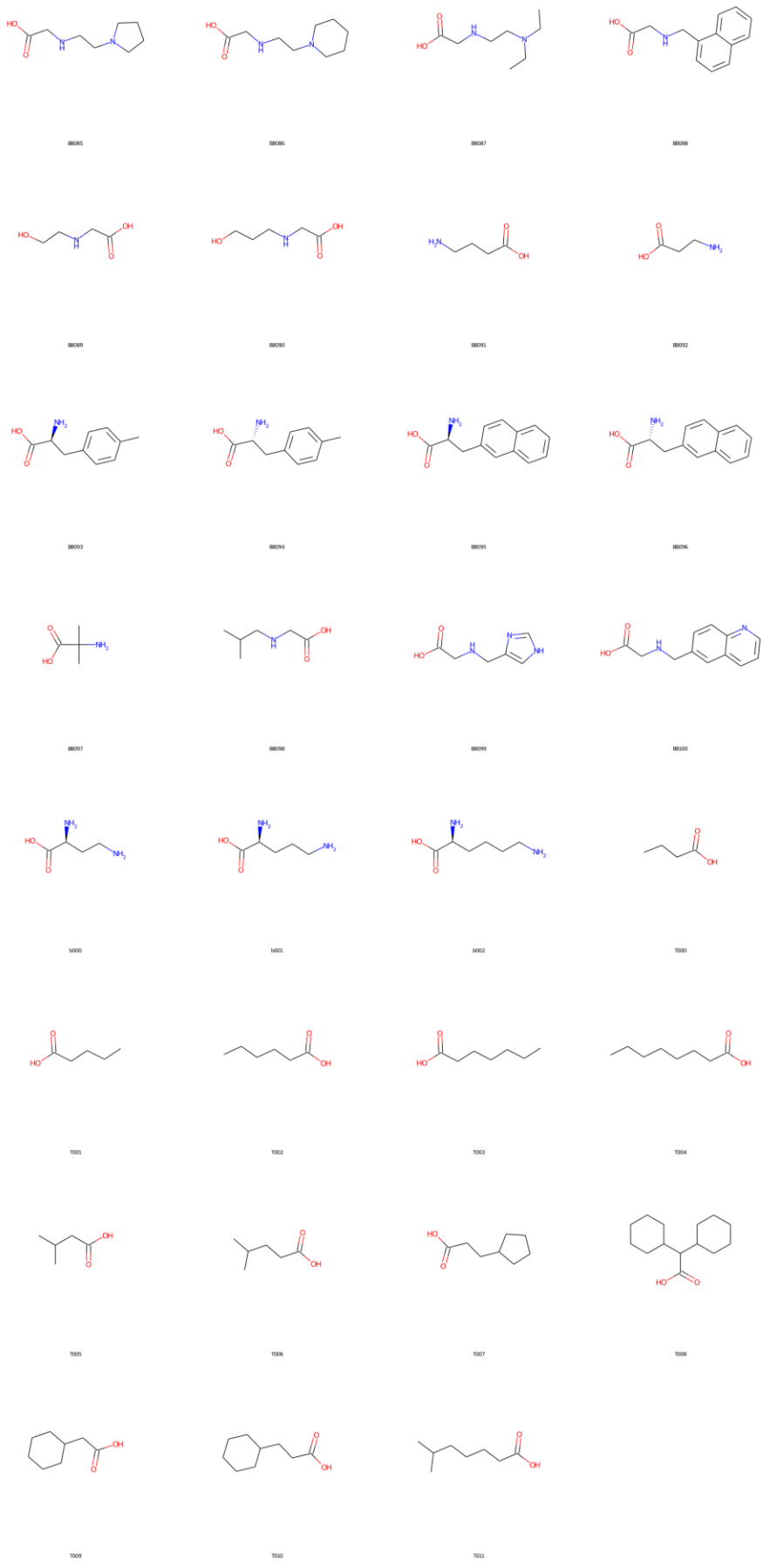


Figure S1: Structures of the building blocks used by the PDGA.

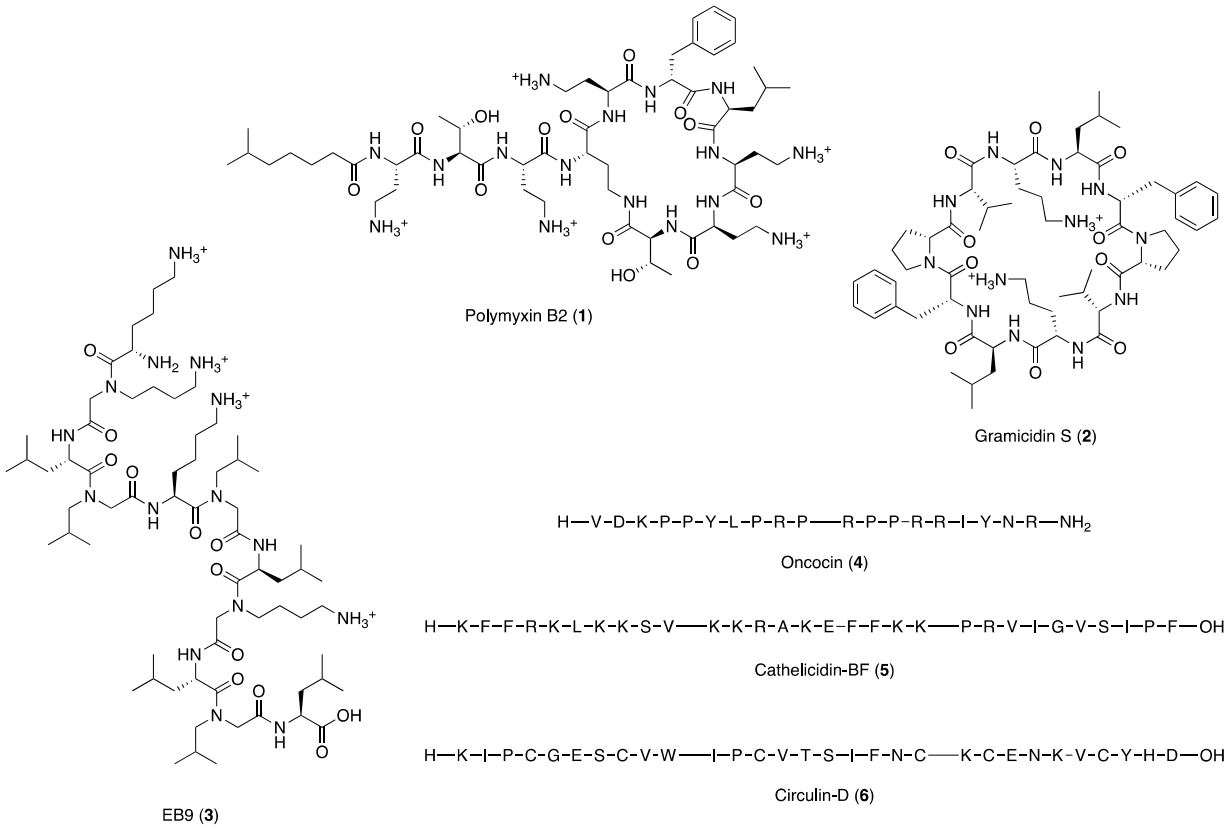


Figure S2: Structures of the selected queries for the PDGA runs using the MAP4C similarity as fitness function.

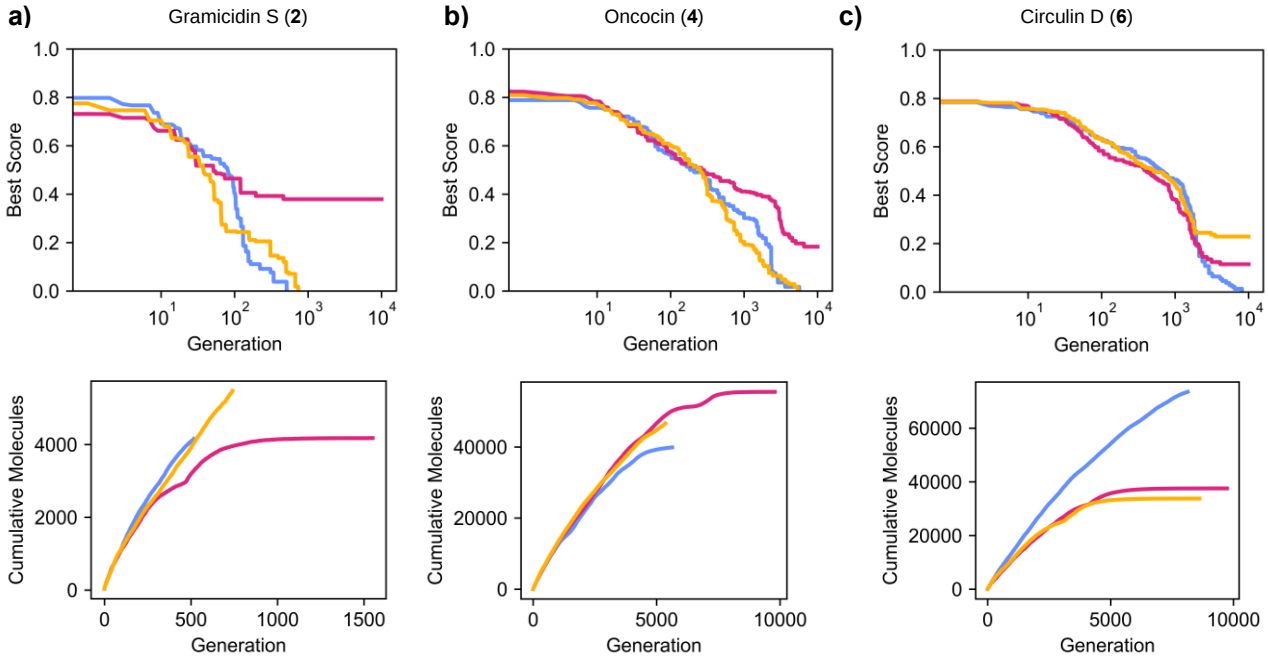


Figure S3: Analysis of three parallel PDGA runs starting from 50 random sequences towards selected queries. Top plots show the overall best score throughout the trajectory; the bottom plots show the cumulative number of unique new molecules generated throughout the trajectory for **a)** gramicidin S, **b)** oncocin, and **c)** circulin D.

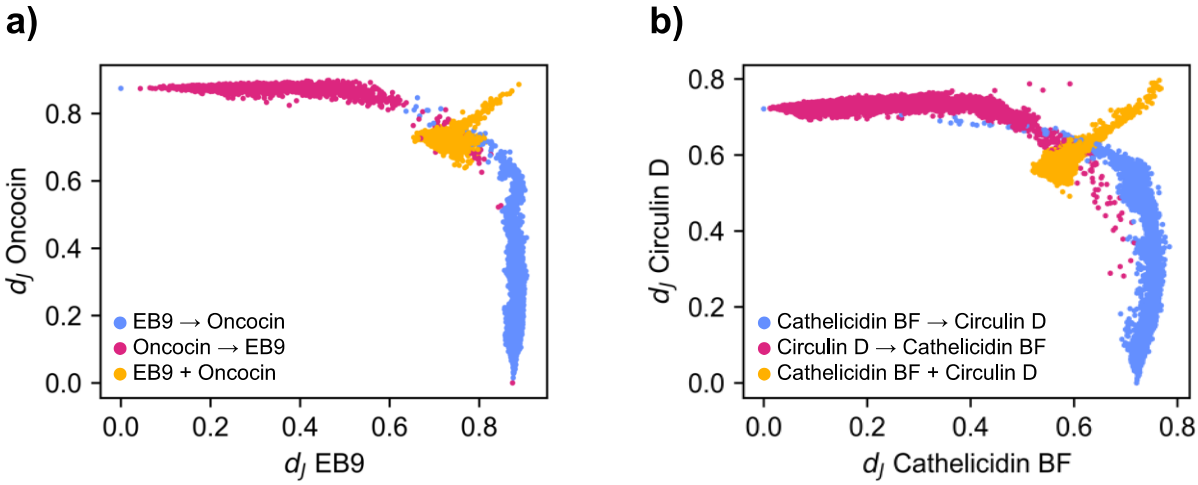


Figure S4: Jaccard distance of molecules selected from the different traversal trajectories towards **a)** oncocin and EB9 and **b)** circulin D and cathelicidin BF.

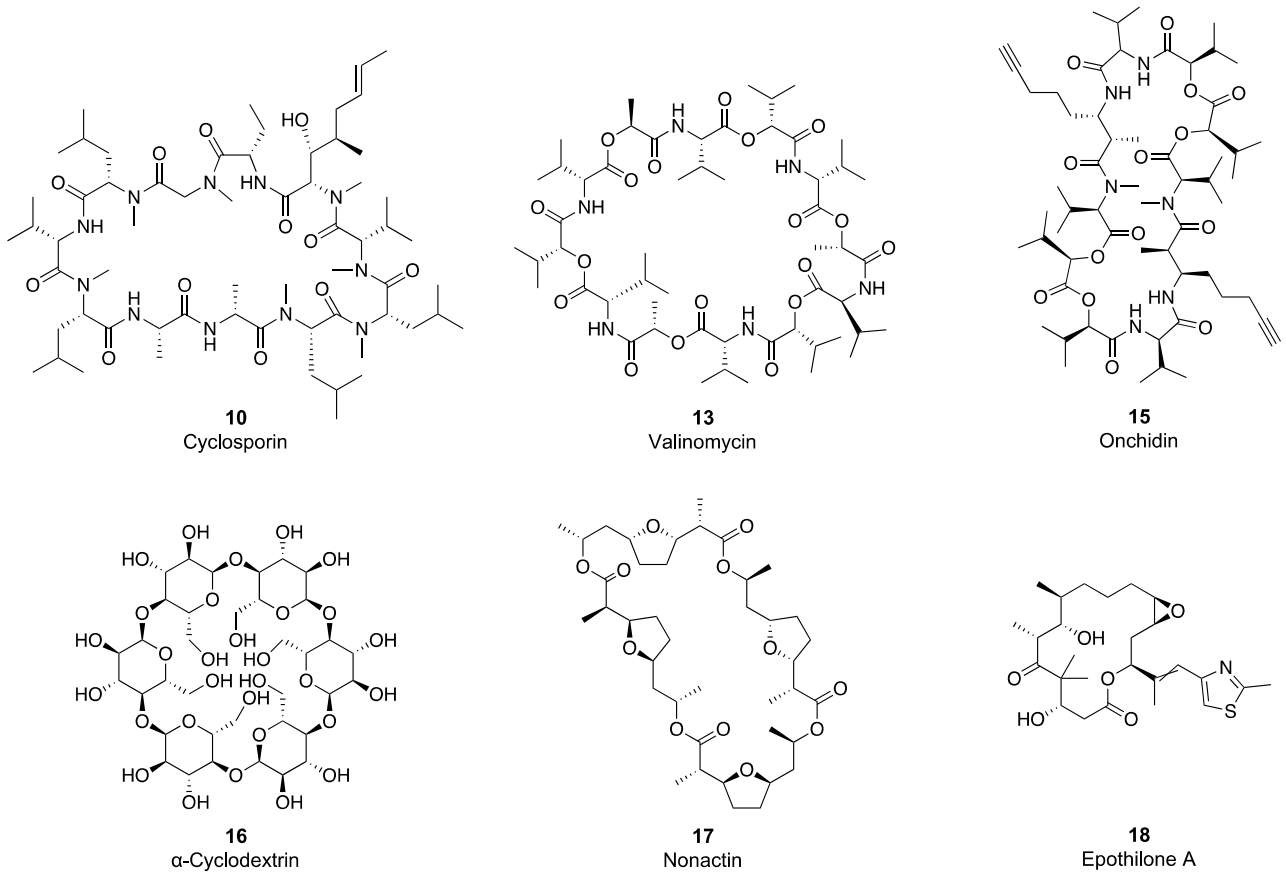


Figure S5: Structures of the non-peptide macrocycle queries for the PDGA runs using the MXFP similarity as fitness function.

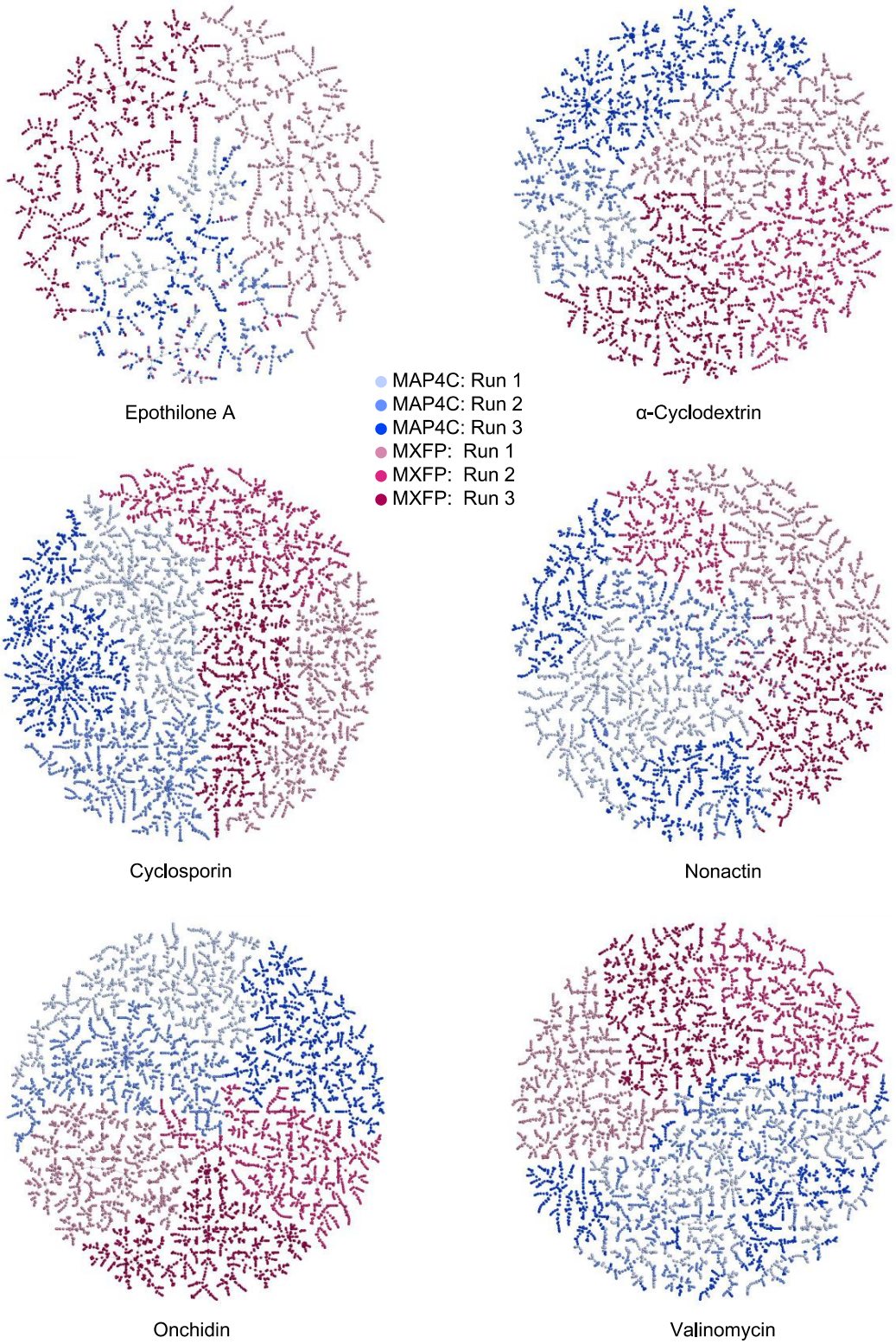


Figure S6: TMAPs of top 1000 molecules generated in each of three parallel MAP4C and MXFP trajectories of selected non-peptide queries. Interactive TMAPs:
https://tm.gdb.tools/map4/10E60/epothilone_tmap.html
https://tm.gdb.tools/map4/10E60/cyclodextrin_tmap.html
https://tm.gdb.tools/map4/10E60/cyclosporin_tmap.html
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