Preprint Cambridge Centre for Computational Chemical Engineering ISSN 1473 – 4273

Leveraging General Molecular Fragments to Expand the Design Space of Metal-Organic Polyhedra

Patrick W.V. Butler[1], Simon D. Rihm[2], Sebastian Mosbach[1,3], Jethro Akroyd[1,3], Markus Kraft[1,2,3]

released: April 17, 2026

[1] Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

[2] CMPG
GRIPS – Gründerinnenzentrum Pirmasens
Delaware Avenue 1–3
66953 Pirmasens
Germany

[3] CARES
Cambridge Centre for Advanced Research and Education in Singapore
1 Create Way
CREATE Tower, #05-05
Singapore, 138602

Preprint No. 345

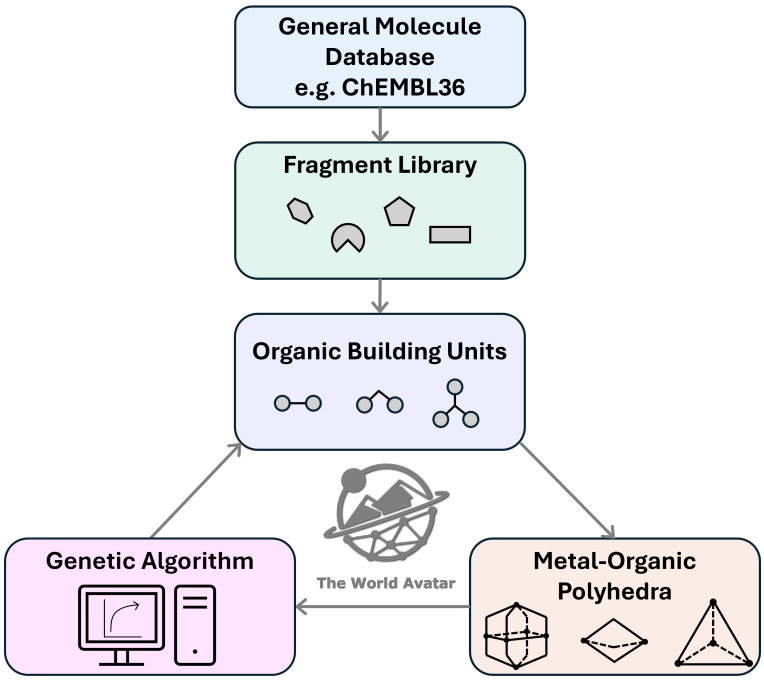
Keywords: Reticular Chemistry, Metal–Organic Polyhedra, Fragment-based Design, Computational Materials Discovery, Evolutionary Optimisation

Edited by
Computational Modelling Group
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom

E-Mail: mk306@cam.ac.uk
World Wide Web: https://como.ceb.cam.ac.uk/

Abstract

The discovery of functional reticular materials is dependent on optimising the organic building unit to achieve precise chemical and structural properties. However, experimentally characterised chemical building units (CBUs) are only a small fraction of the possible design space, and as such, computer-aided design based on datasets of known reticular materials have fundamental limitations in chemical diversity. Here, we present a general workflow for extracting, analysing, and recombining molecular fragments from large, chemically diverse datasets to systematically expand the accessible organic CBU design space. Using the ChEMBL36 database as a case study, we generate a library of 12,387 unique fragments through filtering, fragmentation, and ontology-driven classification. These fragments are recombined to enumerate a dataset of 44.8 million CBUs. To demonstrate integrating the workflow into designing functional reticular materials, we combine the library with a genetic algorithm and optimise for metal-organic polyhedra (MOPs) estimated as stable, non-toxic drug carriers. The concepts and relationships between the molecular fragments, CBUs, templates, and MOPs are captured and connected through a knowledge graph and ontology integrated in The World Avatar.



Highlights

- Workflow to extract molecular fragments suitable for organic building units
- Generate a library of 12,387 fragments from ChEMBL36
- Enumerate a dataset of 44.8 M organic building units over 11 templates
- Using a genetic algorithm, optimise MOPs for a drug delivery application

Contents

1 Introduction

Reticular chemistry offers a powerful framework for the rational design of materials by separating chemical complexity from structural topology through the use of chemical building units (CBUs) and assembly models.[16, 29, 30] While this modularity enables predictable structure formation and systematic property tuning, it also gives rise to an immense combinatorial design space that rapidly exceeds what can be explored experimentally. As a result, computational strategies capable of generating, navigating, and optimising structures are essential for accelerating discovery across reticular materials, including metal-organic frameworks (MOFs), covalent-organic frameworks (COFs), and metal-organic polyhedra (MOPs).[5, 7, 8, 11, 18, 20, 21, 24, 28]

A central focus in developing computational strategies is generating organic CBUs that are both valid building units and sufficiently chemically diverse to meaningfully populate the design space. For validity, this goes beyond normal requirements of valid valency and bonding and includes constraints on binding site geometries and functional groups that enable a CBU to perform the intended role in the structure, and that distinguishes CBUs from arbitrary organic molecules. Two general strategies for exploring the CBU chemical space are: fragment-based methods, in which molecular fragments are recombined according to templates that enforce validity, and data-driven generative models, including recurrent neural networks (RNNs), large-language models (LLMs), and diffusion models, trained to produce plausible building units.[2, 6, 12, 14, 17, 25, 31, 32, 34] Although different in many respects, these two approaches share a fundamental limitation that the chemical diversity of their outputs is dependent on the diversity of the data input, whether a fragment library or a training set. These resources are often curated from known reticular materials, and consequently, large regions of the CBU chemical space remain unexplored, including potentially motifs well-established in other fields of synthetic chemistry.[6, 13, 28] Overcoming this requires either augmenting fragment libraries using chemically diverse sources or training generative models on datasets that extend beyond the known reticular chemistry. While there has been strong interest in the latter approach, the former has been relatively unexplored.

The **purpose of this paper** is to present a general workflow for extracting, analysing, and categorising molecular fragments from arbitrary chemical datasets, enabling systematic construction of large, chemically diverse datasets of organic CBUs for reticular materials design.

2 Results and Discussion

The workflow comprises three stages: filtering the molecular dataset to select compounds most likely to produce useful fragments, decomposing these to generate a library of fragments with assigned types, and assembling the fragments into organic CBUs according to defined templates (Figure 1). We demonstrate this using the ChEMBL36 dataset [33], though the workflow is dataset-agnostic. ChEMBL36, which contains more than 2.8 million bioactive compounds reported primarily in the medicinal chemistry literature, was selected due to its large scale, grounding in experimentally characterised molecules, and the

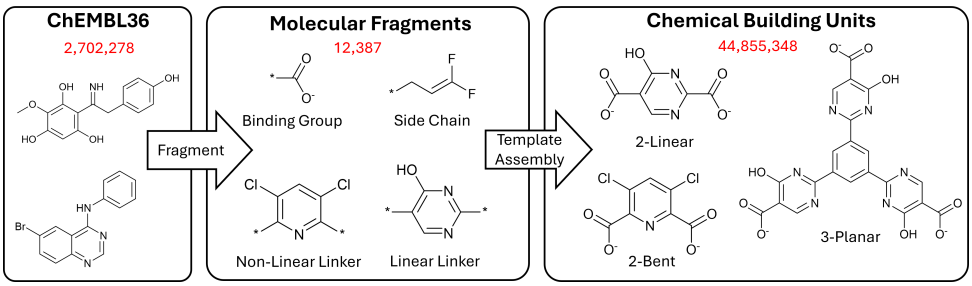


Figure 1: *The workflow for generating CBUs from a general molecule database, in this case ChEMBL36. Following filtering, the molecules are fragmented using the BRICS method and the fragments assigned ontologically defined types based on chemical and geometric analysis. Combining the fragments with CBU templates then allows enumerating large numbers of diverse CBUs.*

likelihood of capturing chemical motifs and functional groups that are under-represented or entirely absent in molecular fragments derived from reticular chemistry.

Applying the workflow, the initial filtering stage selected compounds from ChEMBL36 containing between 10 and 50 non-hydrogen atoms and composed exclusively of elements $X \in \{\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{B}, \mathrm{S}, \mathrm{F}, \mathrm{Cl}, \mathrm{Br}, \mathrm{I}, \mathrm{Si}, \mathrm{P}\}$, following which 2,702,278 compounds remained. These were decomposed using the BRICS method [9], and following analysis and filtering, a library of 12,387 unique fragments was yielded. Using our FragMOPs ontology, the fragments were assigned a type (side chain, binding group, linear linker, non-linear linker, or node) and the data integrated into a knowledge graph accessible through The World Avatar.[3, 6, 19, 22] Analysing these fragments shows that compared to the original FragMOPs set of 95 fragments extracted from reported CBUs, the ChEMBL-derived set has a higher mean molecular weight (177.3 vs. 97.1 g mol^{-1}) and heteroatom fraction (0.29 vs. 0.23). However, the distributions largely overlap (Figure A.2), demonstrating the overall complexity of the new fragment set is comparable to that seen in experimental reticular materials. Using SMARTS matching, we find that 238 functional group motifs are present in the ChEMBL fragments, which is a substantial increase over the 38 found in the FragMOPs set (all 38 of which were also found in the ChEMBL set) and emphasises the high chemical diversity of the ChEMBL fragment library. The new functional groups include some common synthetic motifs, such as secondary alcohols, heteroaromatic oxygen, and lactam groups.

To illustrate the size and diversity of the CBU space accessible with fragments extracted from a general molecular dataset, we enumerated CBUs by recombining the ChEMBL-derived fragments according to a set of CBU templates. These templates were defined based on patterns in reported CBUs and cover 2-linear, 2-bent, and 3-planar geometries. The enumeration was performed using SMILES operations and under a symmetry constraint that ensures the binding sites are equivalent—thus making the enumeration more tractable and avoiding CBUs that would create synthetically challenging chiral MOPs. From a subset of 5924 fragments, selected to further improve synthesisability and tractability (see section A.1.2), a dataset of 44,855,348 unique CBUs was enumerated. A t-SNE plot of 2048-bit Morgan fingerprints (Figure 2) shows that CBUs sampled from this dataset do indeed explore regions of chemical space that are inaccessible using frag-

ments from reported CBUs. Furthermore, as evidenced by lower mean pairwise Tanimoto similarities (0.17 vs. 0.24, Figure A.8), these new CBUs have a higher chemical diversity. A mean SAScore of 3.75 is below that of a number of experimentally reported CBUs (Figure A.7) and suggests that many of these CBUs are synthetically feasible.

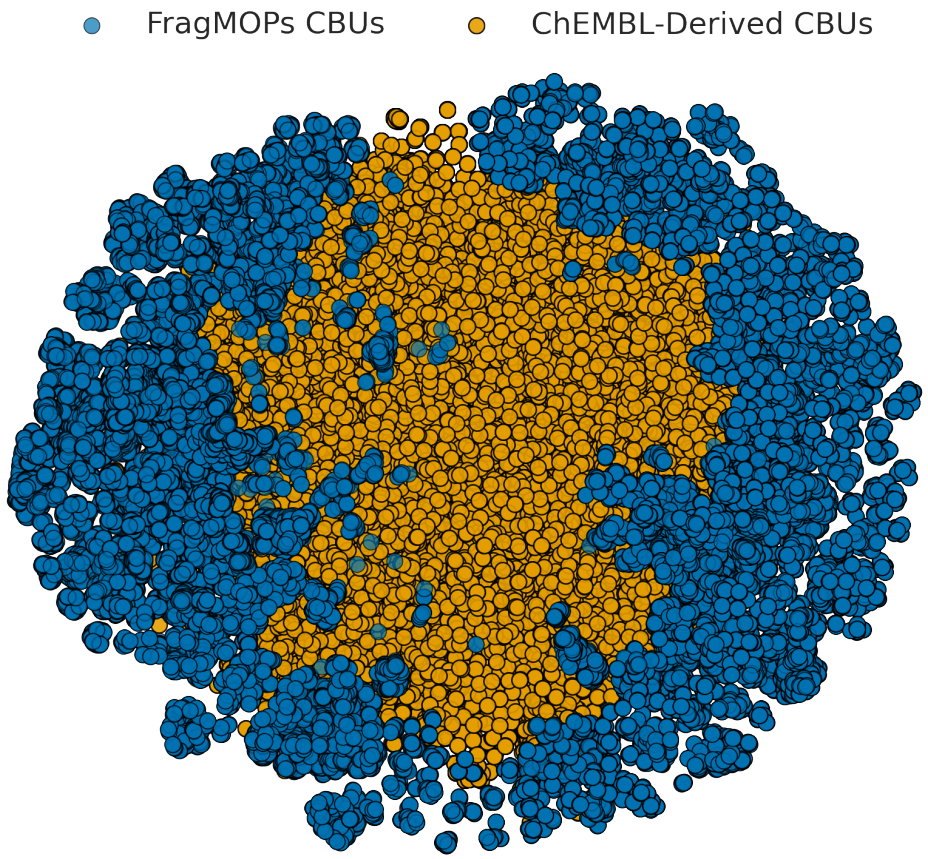


Figure 2: *Depiction of the CBU space as a t-SNE plot calculated using 2048-bit Morgan Fingerprints for the 98,098 organic CBUs assembled from the previous set of reticular chemistry fragments (FragMOPs) and an equal number randomly sampled from the new set derived using the ChEMBL36 fragment library.*

To demonstrate that the expanded organic CBU space can be effectively navigated for applications, we applied a genetic algorithm (GA) to identify (4-planar)$_6$(2-bent)$_{12}$ octahedral MOPs with potential as drug delivery hosts for the antibiotic ciprofloxacin, a payload that has been previously studied for reticular materials to enable controlled and sustained release.[1, 10, 15] The GA optimises MOP properties by evolving a population of chromosomes containing genes encoding the molecular fragments, CBU template, metal CBU, and assembly model, with the fitnesses determined by assembling the encoded MOP and assessing the cavity geometry, CBU toxicity, and synthetic accessibility. The full details of the GA are provided in Section A.1.3. Despite including only the Rh_2 4-planar metal CBU, which is expected to be more likely to be non-toxic, the search space was still estimated to contain nearly 20 million gene combinations.

Across 12 independent trajectories each evolving 100 individuals over 50 generations, the best MOP fitness rapidly increased, being within 20% of the final best fitness after 10 generations. The average of the best MOP fitness for each trajectory shows smoother convergence, increasing consistently over the optimisation. Extracting the top 10 MOPs based on fitness and using the GFN2-xTB method,[4] we calculated an average binding energy for ciprofloxacin following a rigid optimisation of -1.25 eV, indicating stable host-guest complexes. Visualising the optimised structures (Figure 4) finds that the optimised complexes have the carboxylate and ammonium groups of the drug molecule interacting with the Rh sites. Considering there is no experimental MOP comparison available to our knowledge, we compare the fitnesses against that calculated for the MOF UiO-66, which has been studied for ciprofloxacin delivery.[1] The best MOPs exceed the fitness of UiO-66 with the superior performance of the MOPs attributed to their larger cavity diameters (15.3-17.1 Å), which are closer to the target size for readily accommodating ciprofloxacin (16.2 Å) than the largest cavity of UiO-66 (11.0 Å).[27] Despite the successful optimisation, we do not expect that the discovered MOPs will necessarily be high-performing experimentally, due largely to the approximate toxicity and synthesisability models. Rather, these results show that, even under the vastly expanded design space created by the general molecular fragments, a fragment-based optimiser can effectively navigate the space to identify high-performing candidates according to the fitness function.

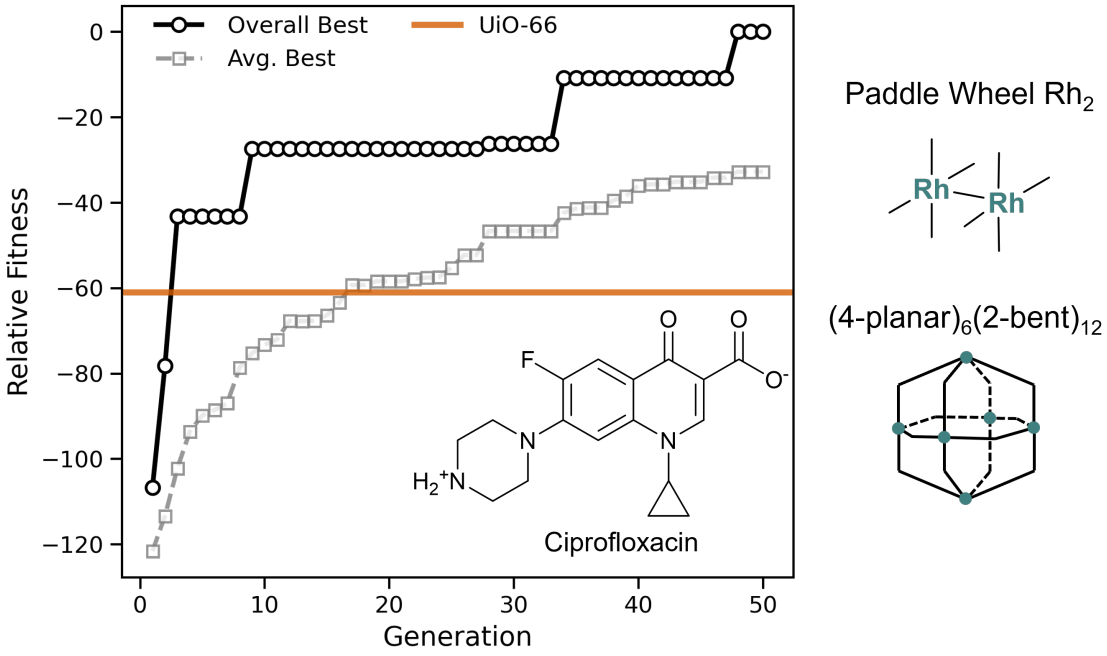


Figure 3: *Convergence of the genetic algorithm optimisation using the ChEMBL fragment library to identify (4-planar)$_6$(2-bent)$_{12}$ MOPs with potential to act as carriers for the antibiotic ciprofloxacin. The best MOP is across all trajectories while the average best is calculated from averaging the best fitness in each trajectory. The relative fitness of the experimental MOF UiO-66 is shown (orange line) as well as depictions of the Rh metal CBU, (4-planar)$_6$(2-bent)$_{12}$ assembly model, and the zwitterionic form of ciprofloxacin.*

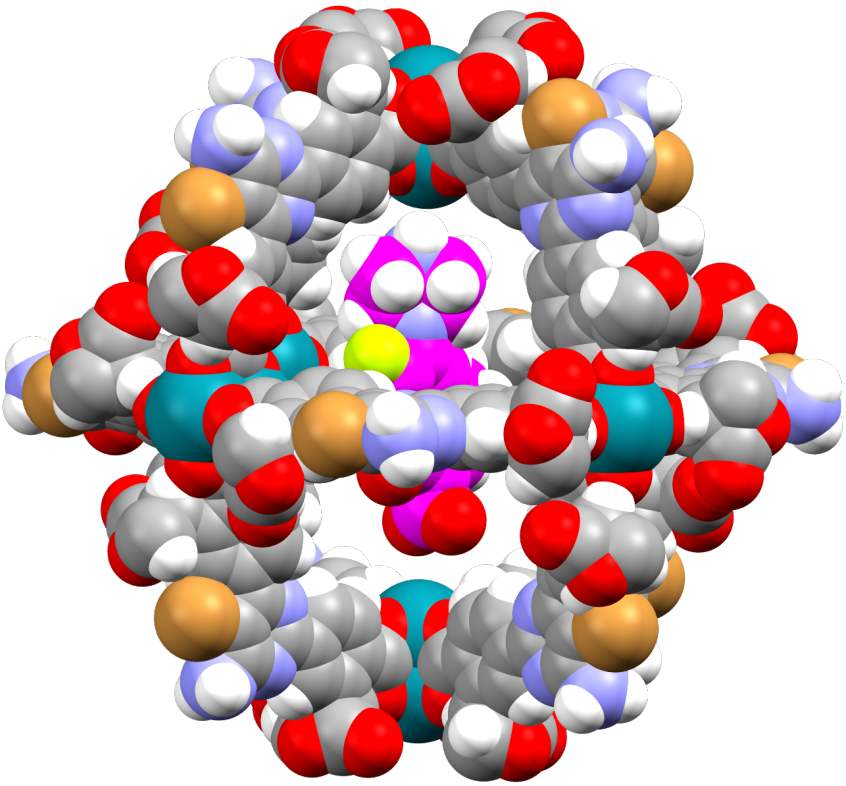


Figure 4: *The most stable host-guest complex identified from the genetic algorithm optimisation targeting MOPs suitable as carriers for ciprofloxacin. The carbon atoms of ciprofloxacin are coloured purple for clarity.*

3 Conclusions

In conclusion, we have presented a workflow that, starting from general molecular data, automatically generates and analyses geometries of fragments to construct organic building units for reticular chemistry. As demonstrated by applying it to the ChEMBL36 data set, with this workflow we were able to create a fragment library of 12,387 fragments. This library captures broad chemical diversity, including functionalities absent from fragments yielded directly from known reticular chemistry, and is ontologically described as part of The World Avatar to support downstream knowledge-driven workflows. Using the fragment library, a dataset of 44.8 million symmetric CBUs was enumerated. Although exhaustive exploration of this space is infeasible, it provides a rich resource for optimising reticular materials for applications, which we demonstrated by integrating in a genetic algorithm optimisation to identify MOPs with potential in drug delivery. These developments provide a scalable workflow to systematically expand and navigate the design space of organic building units, and thereby, accelerate the discovery of high-performing MOPs and related reticular materials.

Acknowledgements

This research was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. This project has received funding from the European Union's Horizon Europe research and innovation programme under grants 101058732 (JIDEP), 101074004 (C2IMPRESS), and 101188248 (CLIMATE-ADAPT4EOSC). This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). M.K. gratefully acknowledges the support of the Alexander von Humboldt Foundation and the Massachusetts Institute of Technology. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Data and code availability

The supplementary information contains additional details on the computational methods, the ontology, CBU template definitions, genetic optimisation results, and examples CBUs and MOPs. The codes and ontologies used are available on GitHub under MIT license: https://github.com/TheWorldAvatar/MOPTools. The data supporting the results is available at Zenodo via DOI: 10.5281/zenodo.19630097.

A Appendix

A.1 Computational Methods

A.1.1 Molecule Fragmentation Workflow

Starting from the SMILES of each molecule, the dataset was filtered to select molecules most likely to yield useful fragments for generating reticular building units. This involved filtering for compounds containing between 10 and 50 non-hydrogen atoms and composed exclusively of elements from the set H, C, N, O, B, S, F, Cl, Br, I, Si, and P. The molecules were then decomposed into fragments using the BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) method.[9] For each fragment, basic descriptors were calculated including the total number of atoms, number of heavy atoms, number of fragmentation (attachment) points, and fraction of non-hydrocarbon atoms. Molecular conformations were then generated using the ETKDGv3 embedding method as implemented in RDKit.[26] For fragments with more than one attachment point a geometric analysis was then performed by placing vectors from the neighbouring atom to each attachment point from which additional properties were calculated including the average and standard deviation of the angle, dihedral, and horizontal offset between attachment points. Based on the properties, the fragments were filtered and categorised according to roles they could fill within CBU templates as defined in our ontology. Firstly, all fragments with a single attachment point were classified as side chains. Fragments with two attachment points were classified as linear linkers if the angle between attachment points exceeded 165°, and as bent linkers if the angle lay between 100° and 130°. Additional filtering criteria were then applied to yield subcategories, for example, the large and small fragments and whether the fragments are aromatic or not.

A.1.2 CBU and MOP Assembly

The CBU assembler takes a CBU template and a set of fragments for each slot in the template. It then orders the fragments based on the positions of each slot and joins them by forming the appropriate bonds. The fragments are always joined starting from a binding group and then adding linker fragments in sequence. For 2-linear and 2-bent templates, after the linkers a second binding group is joined to complete the CBU. For 3-planar templates, copies of the partially assembled structure are added to each bonding site on the node fragment. The direction of asymmetric fragments is determined by a set of binary values that encode which binding site will be bonded first. The assembler supports both SMILES-based methods for rapid enumeration and geometry-based assembly. In geometry-based assembly, local coordinate systems are established for each fragment using bonding site dummy atoms and neighbouring atoms. Fragments are then aligned by transforming coordinate systems to map dummy atoms onto the neighbouring atoms. After removing dummy atoms and forming new bonds, a distance matrix checks for atomic overlaps. If found, the torsion angle increments iteratively until overlaps resolve or it exceeds 360°, at which point the assembly fails. Binding groups are then aligned either to each other for 2-linear templates or to a common plane. A symmetric constraint can

be enabled that enforces asymmetric fragments with multiple positions to have alternating directions. Final MOP generation employs our MOP assembler program that uses geometric operations to position the building units according to a given assembly model.

In the enumeration and later genetic optimisation, for tractability and synthesisability, we limited the fragment library to a subset of 5924 fragments, containing linear linkers with no significant offset between attachment points (< 0.5 Å) and with only carbon atoms neighbouring the fragmentation points.

A.1.3 Genetic Algorithm

For optimising MOP properties with the genetic algorithm, each candidate MOP is encoded by a chromosome with genes that select the components of a MOP including the MOP assembly model, metal CBU and fragments of the organic CBU. An initial population is generated with random genes, and thereafter, subsequent generations are created by ranking the individuals using a defined fitness function that considers properties of the assembled MOP (constructed by the method above) and applying genetic operations to combine chromosomes based on the fitnesses. The two genetic operators used are mutation, which randomly mutates individual genes, and crossover, where genes from two parents in the previous generation are randomly combined to produce new individuals. Additionally, our implementation includes tournament selection, where randomly selected individuals are compared and the winners are used as parents for crossover, as well as elitism, where the fittest individuals are copied directly to the next generation. In this study, we use a population of 100 individuals which are evolved for 50 generations. The elitism percent was 5%, the mutation rate was 0.05, the tournament size was 2, the crossover probability was 0.9, and duplicate genes were removed by applying mutations.

In our previous implementation, we parallelised the genetic algorithm over the CBU templates and used a chromosome containing genes specifying the assembly model, metal CBU, molecular fragments at each position in the template, and the orientations of fragments. While reasonable for smaller search spaces, this approach prevents the algorithm from learning optimal combinations of CBU templates and assembly models and leads to inefficient sampling when chemical spaces associated with different templates significantly vary in size (if sampled with the same number of trajectories). To provide an alternative, we implemented template-aware genetic operators that permit crossover and mutation across templates, allowing for a gene selecting the CBU template to be included in the MOP chromosome. These operators explicitly check whether a proposed fragment is compatible with the allocated position within the current CBU template. If a fragment is not allowed, the mutation is rejected or the corresponding fragment from the parent gene is retained. Additionally, to enable crossover between templates of differing numbers of fragments, all genes are represented with a length equal to that of the largest template, with only those fragment positions relevant to the selected template marked as active and only active genes being considered for crossover.

As a demonstration of the template-aware genetic algorithm, we repeated an optimisation task from our previous study targeting MOPs with a specified inner sphere cavity volume of 625 Å^3. Using the template-aware genetic operators, all optimisation trajectories rapidly converged toward the target cavity volume (Figure A.1). By contrast, using the

previous implementation several CBU templates were unable to reach the target volume due to being locked to small CBU templates, for example, the 2-linear CBU template with one linker fragment. In terms of errors, when parallelising over templates, the maximum deviation from the volume target among the best solutions from each trajectory exceeded 100 Å^3. However, using the template-aware operators, this maximum deviation was reduced to 16 Å^3.

The fitness function used for optimising hosts for ciprofloxacin contained four terms (Equation A.1). The primary geometric target was the internal cavity radius required to optimally accommodate ciprofloxacin, which was chosen over the van der Waals volume due to the molecule being largely planar. Specifically, we targeted a maximum cavity radius of 1.1 times the bounding sphere radius of the zwitterionic form of ciprofloxacin (8.08 Å), this form being favoured under neutral pH conditions. To ensure the GA prioritised geometric fit over secondary descriptors, a high scalar penalty of 20 per Å was applied to deviations from this target. The fitness function further included a prediction of the toxicity of the organic CBU assembled from the fragments using a model recently reported for MOFs. Based on a physiochemical descriptor calculated from the SMILES string, the model returns a classification of fatal, toxic, or safe, and thus these were penalised appropriately with safe CBUs incurring no penalty while fatal received a high penalty (1000) and toxic received half this (500). The magnitude of these toxicity penalties was chosen to be an order of magnitude larger than the other terms, effectively acting as a constraint that directs the search away from hazardous chemical spaces regardless of geometric performance without creating cliffs between safe and fatal/toxic CBUs. The remaining terms of the fitness function were the synthetic accessibility score (SAScore) and a penalty for the number of elements (n_{elem}). Because the SAScore typically ranges from 1 to 10 and n_{elem} is a small integer, these terms were weighted by scalars of 20 and 10, respectively. This scaling ensures that these synthetic considerations contribute meaningfully to the total fitness score—comparable in magnitude to a 1 Å deviation in cavity radius—thereby encouraging the optimiser to explore simpler and more synthesisable CBUs without compromising the primary design objectives. Empirical testing found this weighting gave satisfactory performance for demonstrating the optimisation, allowing for diverse MOPs with cavity properties close to the target.

$$\text{Fitness} = T_{\text{tox}} - 20\,|r_{\text{target}} - r_{\text{cavity}}| - 20 \cdot \text{SAScore} - 10 \cdot n_{\text{elem}} \tag{A.1}$$

where the toxicity penalty T_{tox} is defined as:

$$T_{\text{tox}} = \begin{cases} 0 & \text{if CBU classification is Safe} \\ 500 & \text{if CBU classification is Toxic} \\ 1000 & \text{if CBU classification is Fatal} \end{cases} \tag{A.2}$$

The CBU toxicity classification is predicted using a physicochemical descriptor model based on the SMILES string representation of the organic linker.[23]

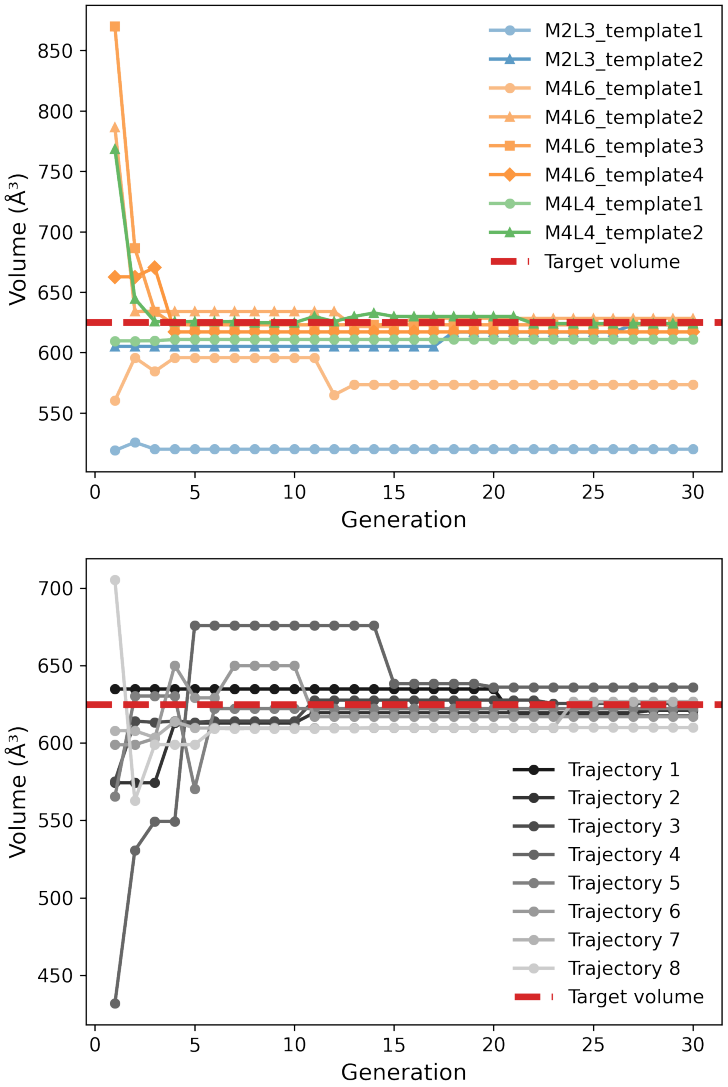


Figure A.1: *The results of the genetic algorithm for optimising MOPs with a maximum inner sphere volume of 625 Å^3 by parallelising over CBU templates (above) and using template-aware genetic operators (below). Both simulations included a total of 8 trajectories.*

A.1.4 Host-Guest Optimisation

To calculate binding energies a rigid-body optimisation was performed between the host MOP and guest molecule. This involved placing the guest molecule at the centre of the MOP and then optimising the orientation of the guest by minimising the energy of the complex as calculated by GFN2-xTB.[4] Once completed a check was performed to ensure the atoms of the host and guest were not overlapping and then the binding energy was calculated as the difference between the host-guest complex and the sum of the isolated energies of the MOP and guest molecule, also calculated with the GFN2-xTB.

A.2 CBU Templates

Table A.1: *Summary of fragment types and sequence positions for each CBU template defined. With the exception of the 2-linear template 1, cyclic linkers with more than 12 heavy atoms are excluded from the accepted fragments. The position values specify the sequence for joining fragments and is only strictly required for linker slots, binding group and node fragments always being joined at the ends of the sequence. Legend: a = aromatic, n = non-aromatic, l = linear, b = bent.*

Template	**Slot 1**		**Slot 2**		**Slot 3**	
	Fragment type	**Pos.**	**Fragment type**	**Pos.**	**Fragment type**	**Pos.**
2-linear template 1	binding group	1, 3	linker (l, a/n)	2	–	–
2-linear template 2	binding group	1, 4	linker (l, a)	2	linear linker (l, a)	3
2-linear template 3	binding group	1, 5	linker (l, a)	2, 4	linker (l, a/n)	3
2-linear template 4	binding group	1, 4	linker (l, n)	2, 4	linker (l, a)	3
2-bent template 1 (acyclic)	binding group	1, 5	linker (l, a)	2, 4	linker (b, n)	3
2-bent template 2 (acyclic)	binding group	1, 7	linker (l, a)	2,3,5 6	linker (b, n)	4
3-planar template 1	binding group	1	linker (l, a)	2	node (3)	3
3-planar template 2	binding group	1	linker (l, a)	2, 3	node (3)	4
2-bent template 1 (cyclic)	binding group	1, 3	linker (b, a)	2	–	–
2-bent template 2 (cyclic)	binding group	1, 5	linker (l, a/n)	2, 4	linker (b, a)	3
2-bent template 3 (cyclic)	binding group	1, 7	linker (b, a)	2, 4	linker (l, a/n)	3

Table A.2: *Summary of the number of unique CBUs assembled for each template.*

Template	**Number of CBUs**
2-linear template 1	5,512
2-linear template 2	7,906
2-linear template 3	16,025,462
2-linear template 4	128,448
2-bent template 1 (acyclic)	3,850,222
2-bent template 2 (acyclic)	3,850,222
3-planar template 1	8,604
3-planar template 2	16,952
2-bent template 1 (cyclic)	2,748
2-bent template 2 (cyclic)	10,955,716
2-bent template 3 (cyclic)	10,005,272
Total	44,855,348

A.3 Fragment Property Distributions

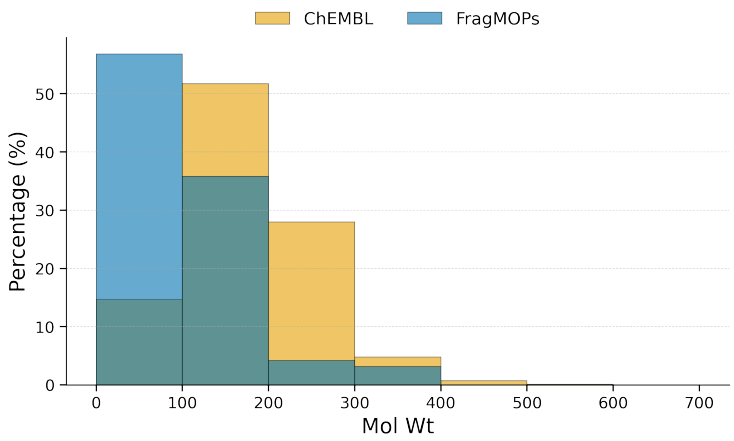


Figure A.2: *Comparison of the molecular weight across the previous FragMOPs set and the new ChEMBL-derived set.*

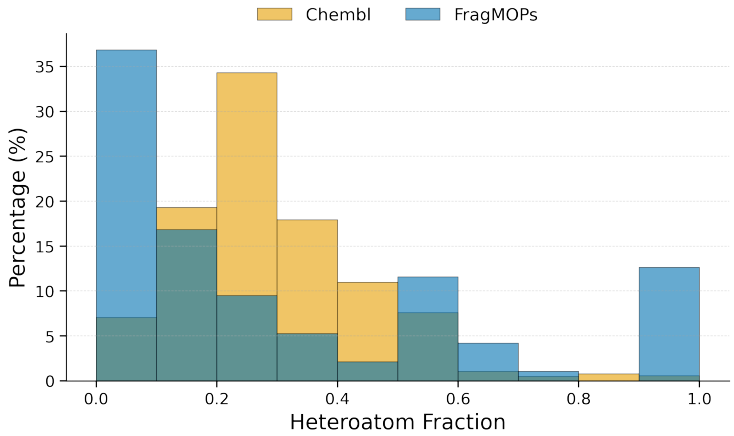


Figure A.3: *Comparison of the heteroatom fraction across the previous FragMOPs set and the new ChEMBL-derived set.*

A.4 CBU Property Distributions

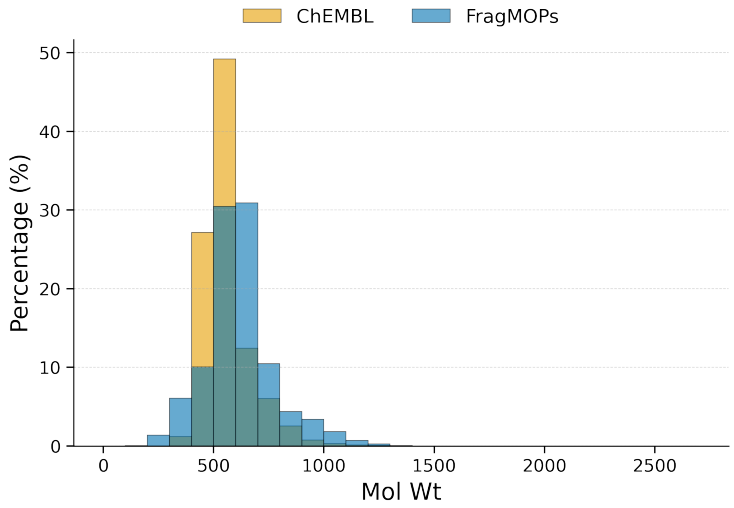


Figure A.4: *Comparison of the molecular weight across the previous FragMOPs set and the new ChEMBL-derived set.*

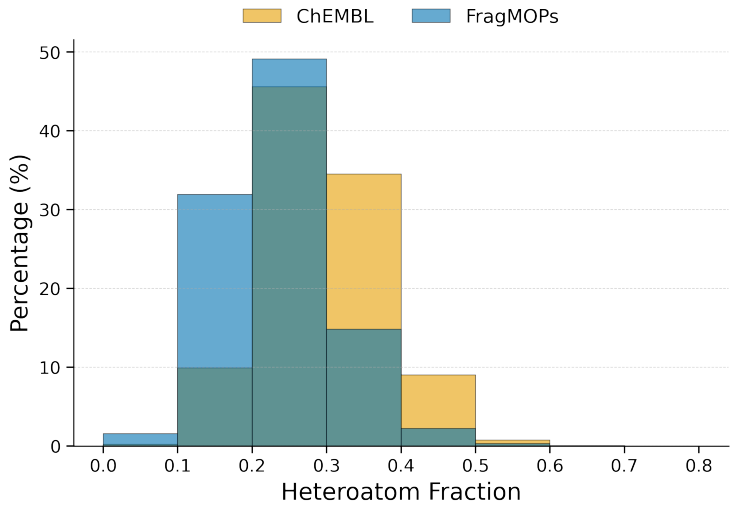


Figure A.5: *Comparison of the heteroatom fraction across the previous FragMOPs set and the new ChEMBL-derived set.*

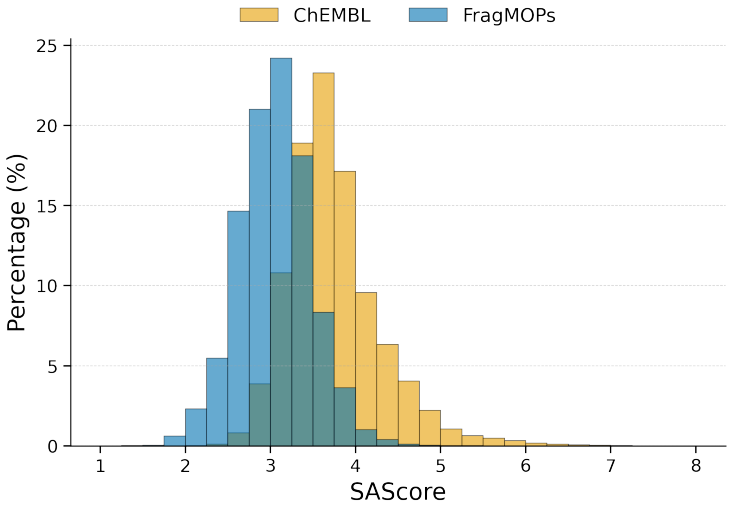


Figure A.6: *Comparison of the SAScore across the previous FragMOPs set and the new ChEMBL-derived set.*

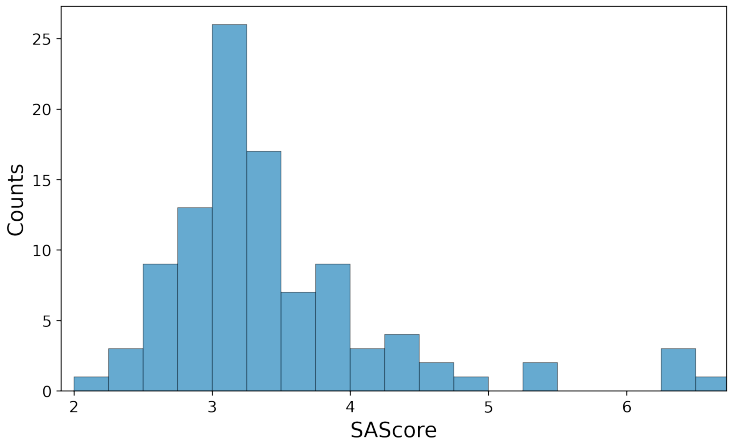


Figure A.7: *SAScore distribution of experimentally characterised organic CBUs from the OntoMOPs dataset.[3, 19]*

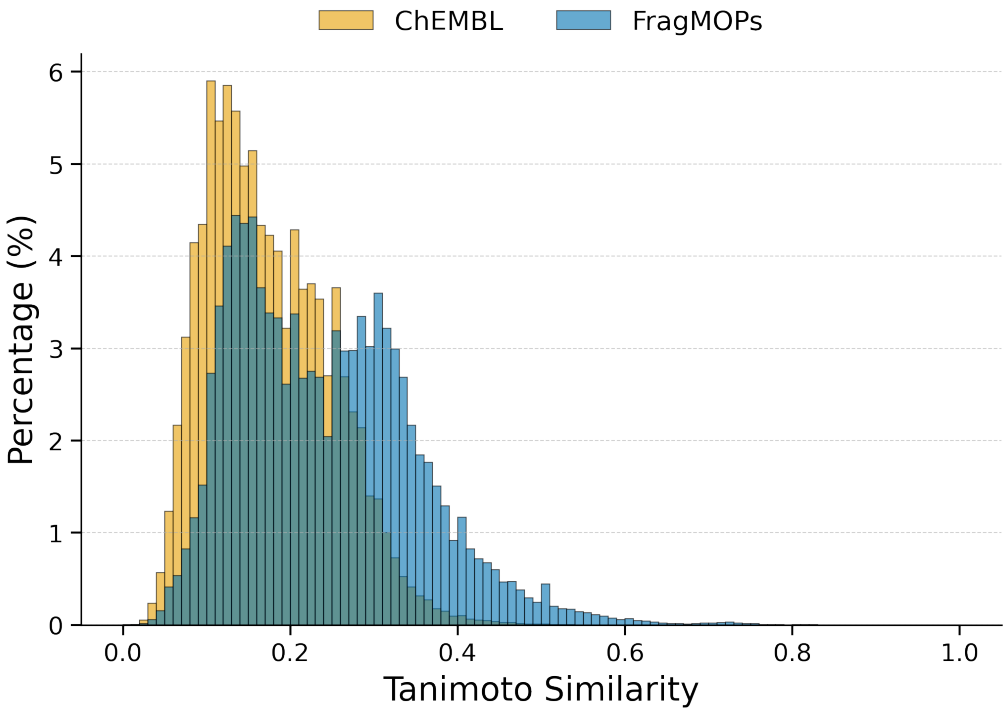


Figure A.8: *Comparison of Tanimoto similarities calculated pairwise using 2048-bit Morgan Fingerprints for the 98,098 organic CBUs assembled from the previous set of reticular chemistry fragments (FragMOPs) and an equal number randomly sampled from the new set derived using the ChEMBL36 fragment library.*

A.5 Example CBUs

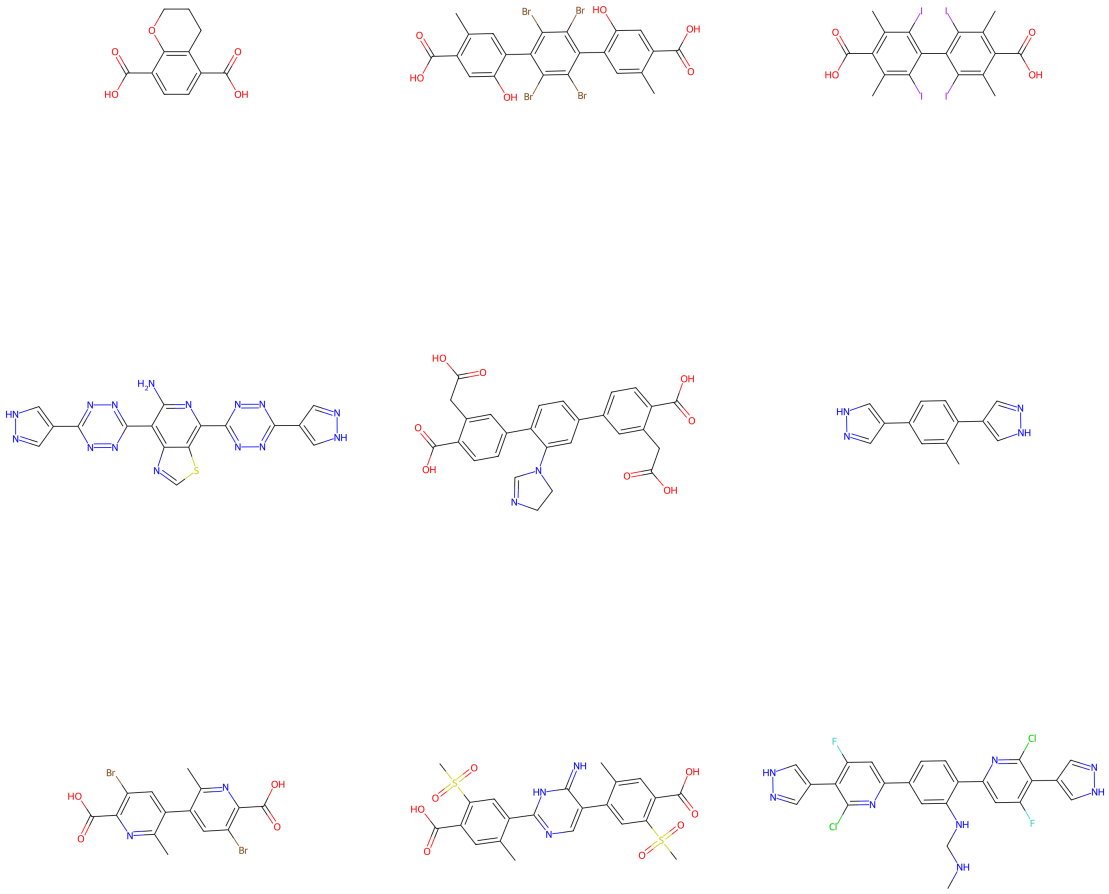


Figure A.9: *Examples of 2-linear CBUs constructed from ChEMBL molecular fragments.*

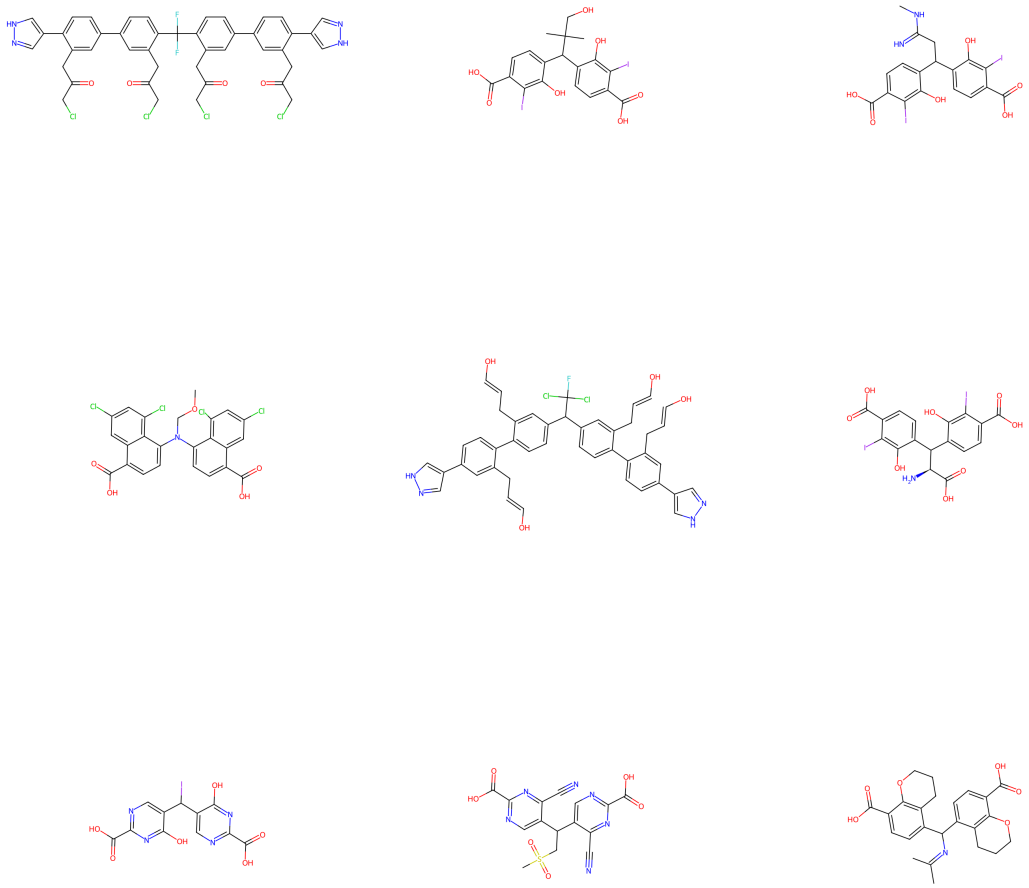


Figure A.10: *Examples of 2-bent CBUs with acyclic non-linear units constructed from ChEMBL molecular fragments.*

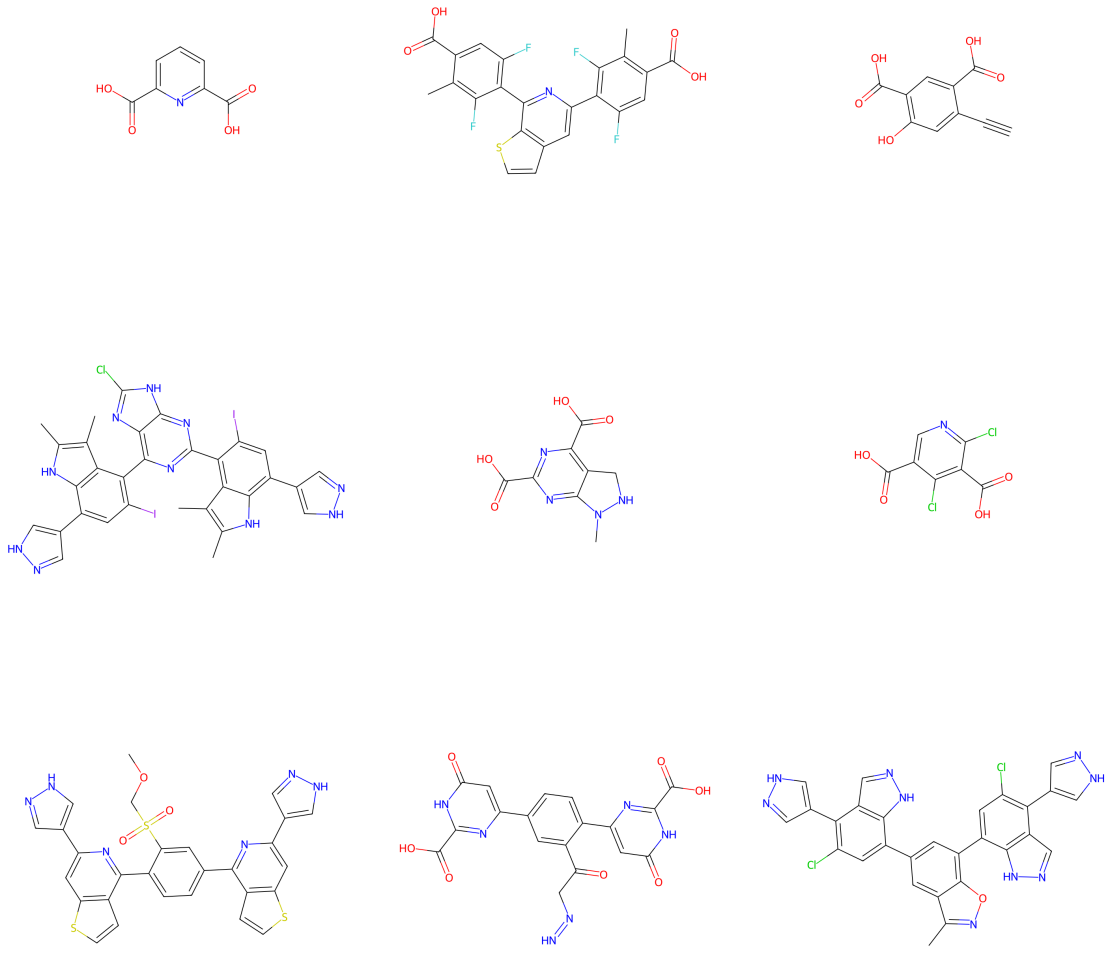


Figure A.11: *Examples of 2-bent CBUs with cyclic non-linear units constructed from ChEMBL molecular fragments.*

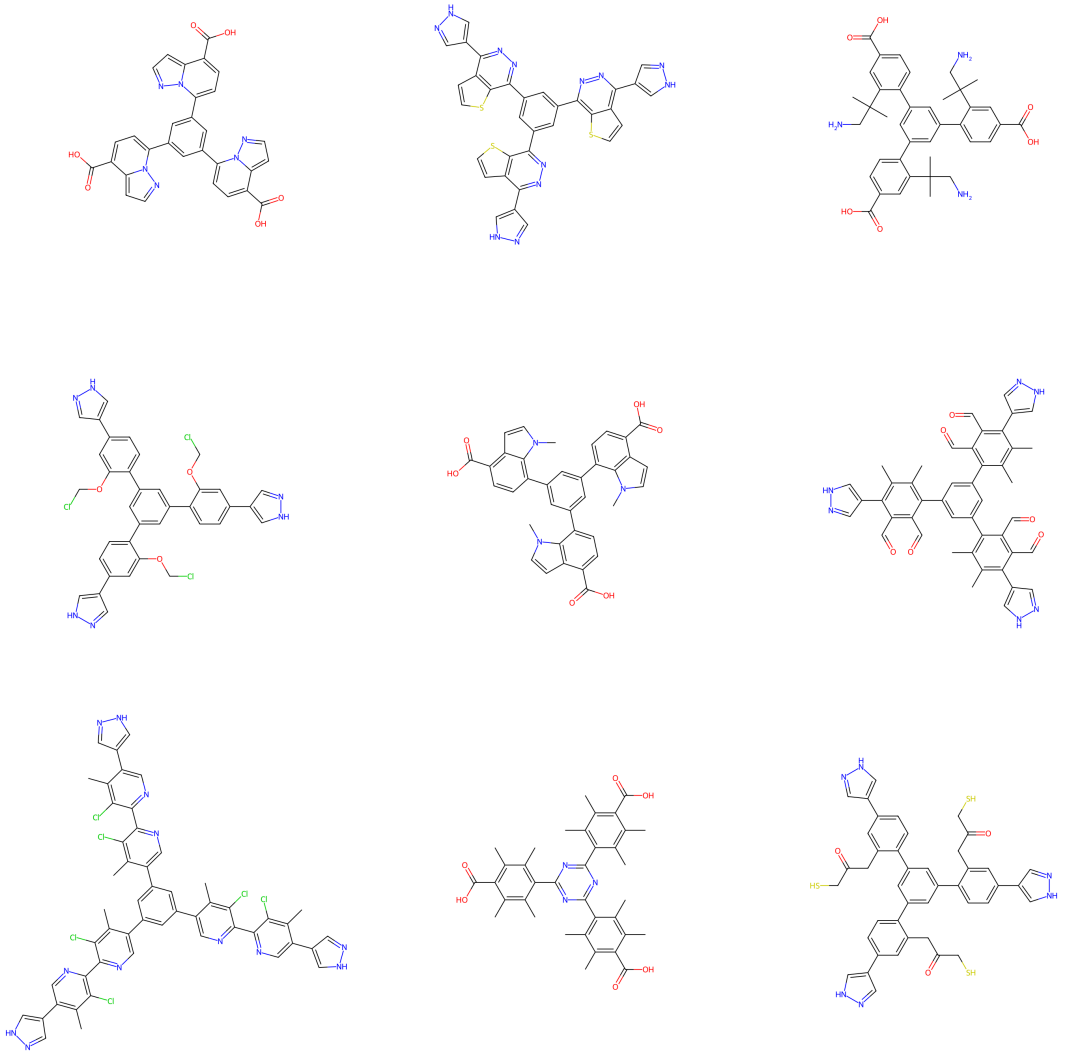
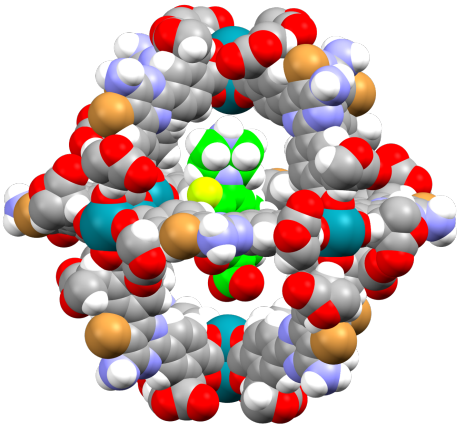
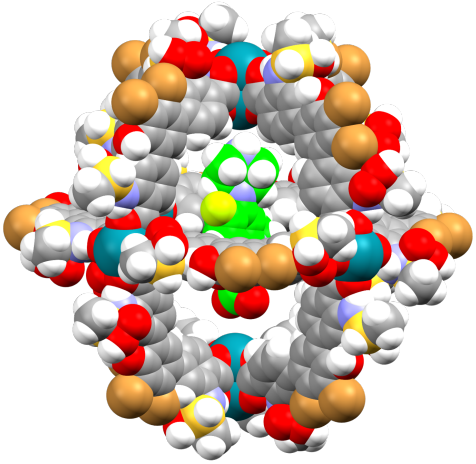


Figure A.12: *Examples of 3-planar CBUs constructed from ChEMBL molecular fragments.*

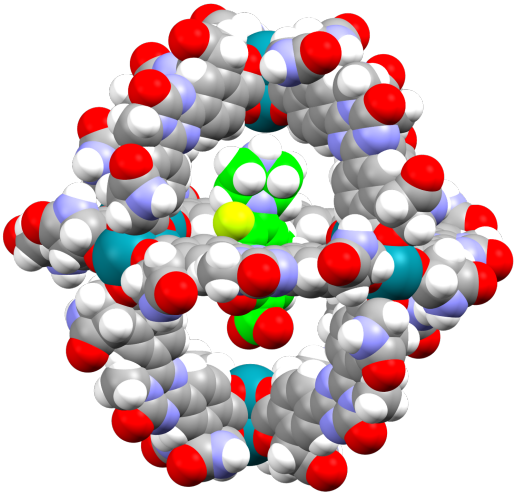
A.6 Genetic Algorithm MOPs



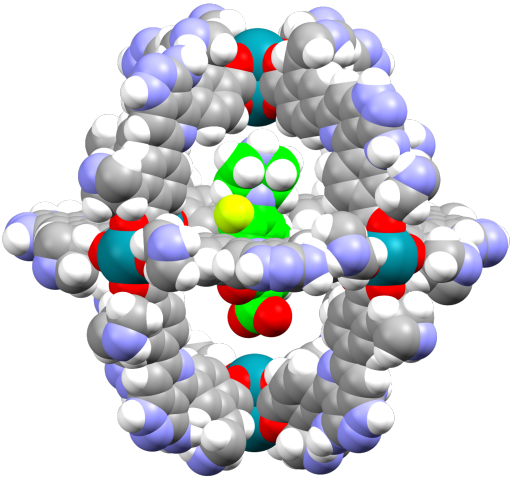
E_{bind} = -1.4916 eV



E_{bind} = -1.1021 eV

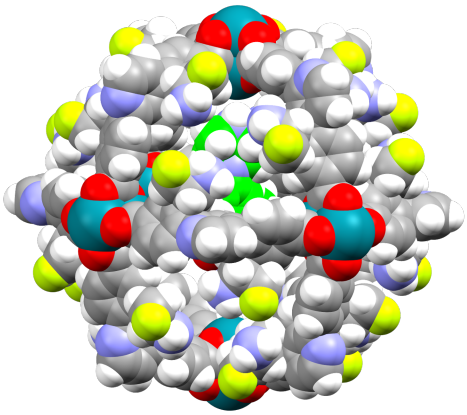


E_{bind} = -1.4602 eV

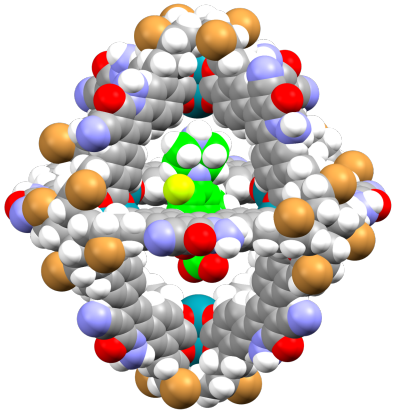


E_{bind} = -1.2081 eV

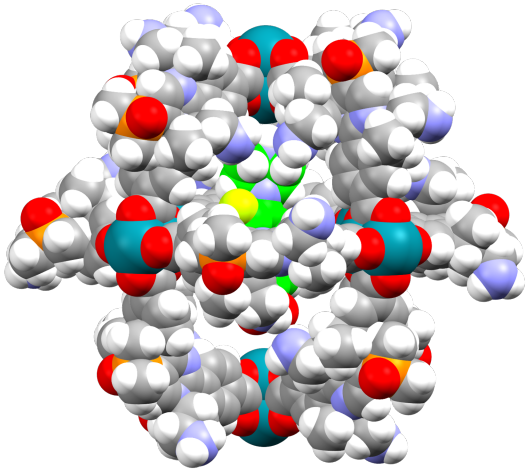
Figure A.13: *Top MOPs from the genetic algorithm optimisation for ciprofloxacin hosts based on fitness. The rigid-body optimised structures and binding energies are shown. MOPs without binding energies were failed due to overlaps between guest and host atoms. Ciprofloxacin carbon atoms have been coloured green for clarity.*



E_{bind} = -0.9966 eV



E_{bind} = -1.3015 eV



E_{bind} = -1.1924 eV

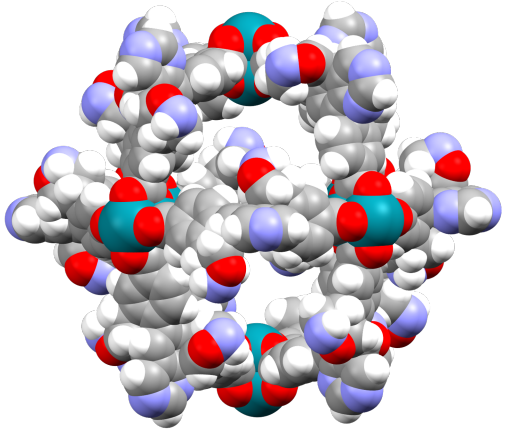


Figure A.14: *cont. Top MOPs from the genetic algorithm optimisation for ciprofloxacin hosts based on fitness. The rigid-body optimised structures and binding energies are shown. MOPs without binding energies were failed due to overlaps between guest and host atoms. Ciprofloxacin carbon atoms have been coloured green for clarity.*

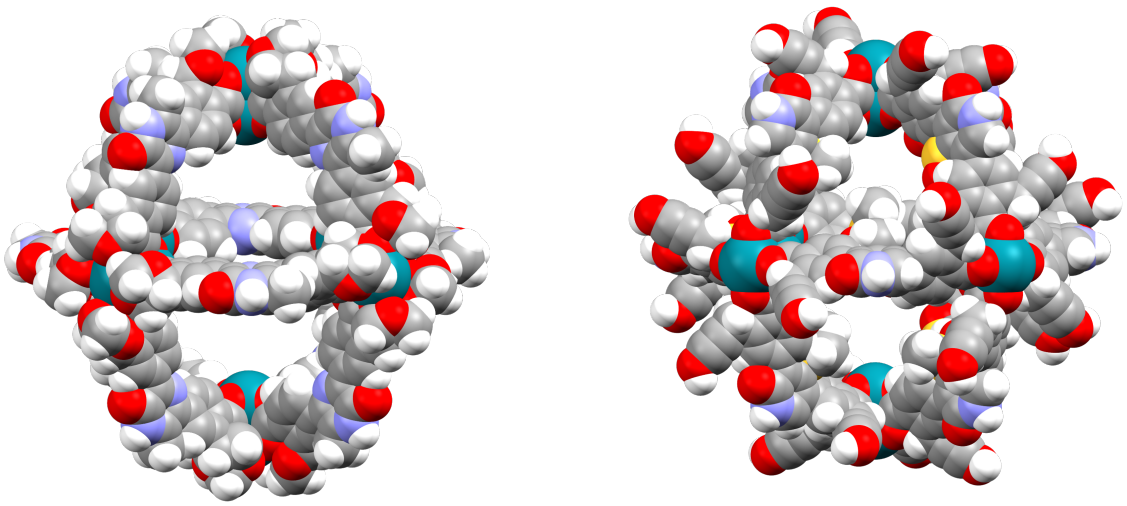


Figure A.15: *cont. Top MOPs from the genetic algorithm optimisation for ciprofloxacin hosts based on fitness. The rigid-body optimised structures and binding energies are shown. MOPs without binding energies were failed due to overlaps between guest and host atoms. Ciprofloxacin carbon atoms have been coloured green for clarity.*

References

[1] S. F. Aden, L. A. M. Mahmoud, E. H. Ivanovska, L. R. Terry, V. P. Ting, M. G. Katsikogianni, and S. Nayak. Controlled delivery of ciprofloxacin using zirconium-based MOFs and poly-caprolactone composites. *Journal of Drug Delivery Science and Technology*, 88:104894, 2023. doi:10.1016/j.jddst.2023.104894.

[2] S. Badrinarayanan, R. Magar, A. Antony, R. S. Meda, and A. Barati Farimani. MOFGPT: Generative Design of Metal–Organic Frameworks using Language Models. *Journal of Chemical Information and Modeling*, 65(17):9049–9060, 2025. doi:10.1021/acs.jcim.5c01625.

[3] J. Bai, S. D. Rihm, A. Kondinski, F. Saluz, X. Deng, G. Brownbridge, S. Mosbach, J. Akroyd, and M. Kraft. Twa: The World Avatar Python package for dynamic knowledge graphs and its application in reticular chemistry. *Digital Discovery*, 4(8): 2123–2135, 2025. doi:10.1039/D5DD00069F.

[4] C. Bannwarth, S. Ehlert, and S. Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. doi:10.1021/acs.jctc.8b01176.

[5] P. G. Boyd and T. K. Woo. A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm*, 18 (21):3777–3792, 2016. doi:10.1039/C6CE00407E.

[6] P. W. Butler, S. D. Rihm, S. Mosbach, J. Akroyd, and M. Kraft. Inverse Design of Metal–Organic Polyhedra through Molecular Fragmentation and Evolutionary Optimisation. *Journal of Chemical Information and Modeling*, page acs.jcim.5c02956, 2026. doi:10.1021/acs.jcim.5c02956.

[7] Y. J. Colón, D. A. Gómez-Gualdrón, and R. Q. Snurr. Topologically guided, automated construction of metal–organic frameworks and their evaluation for energy-related applications. *Crystal Growth & Design*, 17(11):5801–5810, 2017. doi:10.1021/acs.cgd.7b00848.

[8] K. S. Deeg, D. Damasceno Borges, D. Ongari, N. Rampal, L. Talirz, A. V. Yakutovich, J. M. Huck, and B. Smit. In Silico Discovery of Covalent Organic Frameworks for Carbon Capture. *ACS Applied Materials & Interfaces*, 12(19):21559–21568, 2020. doi:10.1021/acsami.0c01659.

[9] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi:10.1002/cmdc.200800178.

[10] A. Domke, M. Fischer, M. Jakubowski, A. Pacholak, M. Ratajczak, A. Voelkel, and M. Sandomierski. Experimental and computational study on the Ca2+, Mg2+, Zn2+ and Sr2+ exchanged zeolites as a drug delivery system for fluoroquinolone

antibiotic – Ciprofloxacin. *Journal of Drug Delivery Science and Technology*, 99: 105997, 2024. doi:10.1016/j.jddst.2024.105997.

[11] C. Duan, A. Nandy, S. C. Pal, X. Yang, W. Gao, Y. Du, H. Kraß, Y. Kang, V. Bernales, Z. Ye, T. Pyle, R. Yang, Z. Gu, P. Schwaller, S. Ma, S. Sun, A. Aspuru-Guzik, S. M. Moosavi, R. Wexler, and Z. Zheng. The rise of generative AI for metal-organic framework design and synthesis. *Matter*, 0(0), 2026. doi:10.1016/j.matt.2026.102748.

[12] X. Fu, T. Xie, A. S. Rosen, T. Jaakkola, and J. Smith. MOFDiff: Coarse-grained Diffusion for Metal-Organic Framework Design. (arXiv:2310.10732), 2023. doi:10.48550/arXiv.2310.10732.

[13] D. A. Gómez-Gualdrón, Y. J. Colón, X. Zhang, T. C. Wang, Y.-S. Chen, J. T. Hupp, T. Yildirim, O. K. Farha, J. Zhang, and R. Q. Snurr. Evaluating topologically diverse metal–organic frameworks for cryo-adsorbed hydrogen storage. *Energy & Environmental Science*, 9(10):3279–3289, 2016. doi:10.1039/C6EE02104B.

[14] T. J. Inizan, S. Yang, A. Kaplan, Y.-h. Lin, J. Yin, S. Mirzaei, M. Abdelgaid, A. H. Alawadhi, K. Cho, Z. Zheng, E. D. Cubuk, C. Borgs, J. T. Chayes, K. A. Persson, and O. M. Yaghi. System of Agentic AI for the Discovery of Metal-Organic Frameworks. (arXiv:2504.14110), 2025. doi:10.48550/arXiv.2504.14110.

[15] M. Ishfaq, D. Lateef, Z. Ashraf, M. Sajjad, M. Owais, W. Shoukat, M. Mohsin, M. Ibrahim, F. Verpoort, and A. Hussain Chughtai. Zirconium-based MOFs as pH-responsive drug delivery systems: Encapsulation and release profiles of ciprofloxacin. *RSC Advances*, 15(33):26647–26659, 2025. doi:10.1039/D5RA01665G.

[16] H. Jiang, D. Alezi, and M. Eddaoudi. A reticular chemistry guide for the design of periodic solids. *Nature Reviews Materials*, 6(6):466–487, 2021. doi:10.1038/s41578-021-00287-y.

[17] Y. Kang and J. Kim. ChatMOF: An artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature communications*, 15(1):4705, 2024. doi:10.1038/s41467-024-48998-4.

[18] A. Kondinski, A. Menon, D. Nurkowski, F. Farazi, S. Mosbach, J. Akroyd, and M. Kraft. Automated Rational Design of Metal–Organic Polyhedra. *Journal of the American Chemical Society*, 144(26):11713–11728, 2022. doi:10.1021/jacs.2c03402.

[19] A. Kondinski, A. M. Oyarzún, S. D. Rihm, J. Bai, S. Mosbach, J. Akroyd, and M. Kraft. Automated Assembly Modelling of Metal-Organic Polyhedra. *European Journal of Inorganic Chemistry*, page e202500115, 2025. doi:10.1002/ejic.202500115.

[20] Y. Lan, X. Han, M. Tong, H. Huang, Q. Yang, D. Liu, X. Zhao, and C. Zhong. Materials genomics methods for high-throughput construction of COFs and targeted synthesis. *Nature Communications*, 9(1):5274, 2018. doi:10.1038/s41467-018-07720-x.

[21] S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho, and J. Kim. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021. doi:10.1021/acsami.1c02471.

[22] M. Q. Lim, X. Wang, O. Inderwildi, and M. Kraft. The World Avatar—A World Model for Facilitating Interoperability. In O. Inderwildi and M. Kraft, editors, *Intelligent Decarbonisation: Can Artificial Intelligence and Cyber-Physical Systems Help Achieve Climate Mitigation Targets?*, pages 39–53. Springer International Publishing, Cham, 2022. ISBN 978-3-030-86215-2. doi:10.1007/978-3-030-86215-2_4.

[23] D. Menon and D. Fairen-Jimenez. Guiding the rational design of biocompatible metal-organic frameworks for drug delivery. *Matter*, 8(3), 2025. doi:10.1016/j.matt.2025.101958.

[24] R. Mercado, R.-S. Fu, A. V. Yakutovich, L. Talirz, M. Haranczyk, and B. Smit. In Silico Design of 2D and 3D Covalent Organic Frameworks for Methane Storage Applications. *Chemistry of Materials*, 30(15):5069–5086, 2018. doi:10.1021/acs.chemmater.8b01425.

[25] J. Park, Y. Lee, and J. Kim. Multi-modal conditional diffusion model using signed distance functions for metal-organic frameworks generation. *Nature Communications*, 16(1):34, 2025. doi:10.1038/s41467-024-55390-9.

[26] S. Riniker and G. A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling*, 55(12):2562–2574, 2015. doi:10.1021/acs.jcim.5b00654.

[27] F. Vermoortele, R. Ameloot, A. Vimont, C. Serre, and D. D. Vos. An amino-modified Zr-terephthalate metal–organic framework as an acid–base catalyst for cross-aldol condensation. *Chemical Communications*, 47(5):1521–1523, 2011. doi:10.1039/C0CC03038D.

[28] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, and R. Q. Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry*, 4(2):83–89, 2012. doi:10.1038/nchem.1192.

[29] S. Wuttke. Toward the Nobel Prize: Dissecting Fundamental Principles and Applications of MOF and COF Materials. *Advanced Materials*, 37(52):e71859, 2025. doi:10.1002/adma.71859.

[30] O. M. Yaghi. Reticular Chemistry—Construction, Properties, and Precision Reactions of Frameworks. *Journal of the American Chemical Society*, 138(48):15507–15509, 2016. doi:10.1021/jacs.6b11821.

[31] Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, and Á. Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021. doi:10.1038/s42256-020-00271-1.

[32] T. A. Young, R. Gheorghe, and F. Duarte. Cgbind: A Python Module and Web App for Automated Metallocage Construction and Host–Guest Characterization. *Journal of Chemical Information and Modeling*, 60(7):3546–3557, 2020. doi:10.1021/acs.jcim.0c00519.

[33] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, and A. R. Leach. The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024. doi:10.1093/nar/gkad1004.

[34] X. Zhang, K. Zhang, and Y. Lee. Machine Learning Enabled Tailor-Made Design of Application-Specific Metal–Organic Frameworks. *ACS Applied Materials & Interfaces*, 12(1):734–743, 2020. doi:10.1021/acsami.9b17867.