Preprint CheMIn Technical Reports ISSN 3061-1075

Geometric Host–Guest Screening of Metal-Organic Polyhedra

Aleksandar Kondinski[1], Leon Asanovski[2], Gorazd Bishkoski[2], Gorjan Bogoevski[2], Dimitar Trajanov[2,3]

released: 29 August 2026

[1] Institute of Physical and Theoretical Chemistry
Graz University of Technology
Stremayrgasse 9, 8010 Graz, Austria

[2] Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje, Macedonia

[3] Department of Computer Science,
Metropolitan College
Boston University
Boston, MA 02215, USA

Preprint No. 7



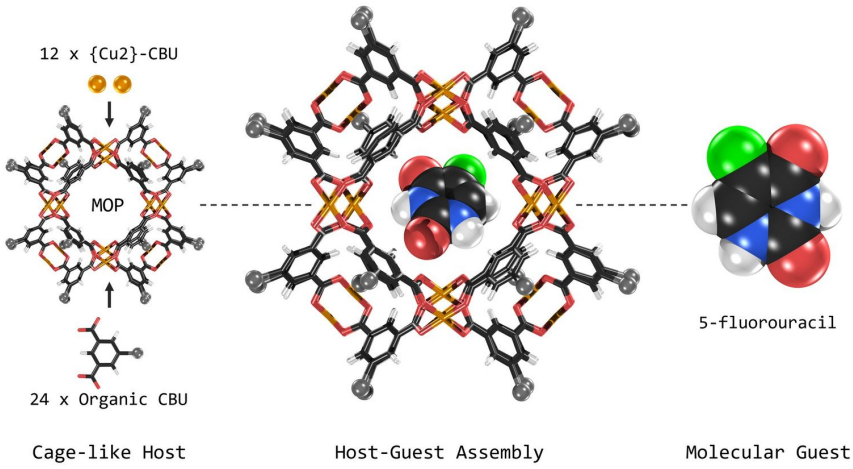
Keywords: geometric screening, host–guest chemistry, metal-organic polyhedra, polyoxometalates, pore engineering

Edited by
Chemical Modelling and Informatics Group
Institute of Physical and Theoretical Chemistry
Graz University of Technology
Stremayrgasse 9, 8010 Graz
Austria (Europe)

E-Mail: kondinski@tugraz.at
World Wide Web: http://chemin.at/

Abstract

Large host–guest libraries of molecular cages cannot be screened experimentally by exhaustive pairwise testing. We present an open molecular-informatics workflow that evaluates 1,471 metal–organic polyhedra against 1,936 polyoxometalates and 292 drug and research compounds, comprising 3.28 million pairwise comparisons. Guest outer diameter was compared with pore-opening diameter for entry and inner-cavity diameter for residence. Pore opening is smaller than the cavity diameter for 99.1% of the hosts and therefore acts as the dominant geometric bottleneck in this library. Among 5,688 POM–MOP candidates with 70–90% cavity filling, 14 hosts from one T_d-symmetric assembly-model topology account for 78% of all candidates (exact one-sided Mann–Whitney U test, $p = 4.1 \times 10^{-9}$, rank-biserial $r = 0.69$), while a second topology contributes a further 17%. The workflow converts molecular datasets into an actionable compatibility map for prioritising simulation and experiment. All data, code, and interactive tools are openly available.



Highlights

- 3.28 million MOP host-guest pairs screened geometrically.
- Pore opening, not cavity size, is the dominant bottleneck.
- One Td-symmetric topology accounts for 78
- Open workflow, data, and interactive web application.

Contents

1 Introduction **3**

2 Methodology **4**

2.1 Data Sources 4

2.2 Molecular Descriptor Calculation 5

2.3 Systematic Screening 6

2.4 Close-Fitting Pair Selection 6

2.5 Statistical Analysis 7

3 Results **7**

3.1 Dataset Overview 7

3.2 Assembly Model Distribution 8

3.3 Compound Dataset Characterisation 8

3.4 POM Screening Results 9

3.5 MOP Host Results 11

3.6 Compatibility Analysis 11

3.7 Close-Fitting Pairs 12

4 Discussion **13**

4.1 Relative Importance of Entry and Cavity Size 13

4.2 Assembly Model Concentration of Close-Fitting Pairs 14

4.3 Interpreting Compound Compatibility and Retrospective Validation . . . 15

5 Conclusions **15**

References **18**

1 Introduction

Host–guest chemistry in molecular cages demonstrates how confinement can alter molecular recognition, reactivity, transport, and guest stability relative to bulk solution [1–4]. These effects arise from the restricted space, defined geometry, and chemical environment of the host cavity. Guest binding can also be controlled through electrostatic interactions, as shown for charged guests encapsulated within $M_{12}L_{24}$ coordination cages [5]. A broad range of molecular hosts has now been developed, including water-soluble coordination cages, Pd- and Pt-based metallosupramolecular cages, covalent molecular cages, and porous organic cages [6–10]. Together, these systems provide a well-established basis for studying how cage structure, cavity dimensions, and internal functionality govern host–guest compatibility.

Metal-organic polyhedra (MOPs) are well suited to database-scale geometric screening because they combine discrete molecular structures with accessible internal cavities and modular construction [11–13]. Recent advances in assembly modelling, geometric descriptor calculation, and knowledge-graph representation now allow systematic comparison across large MOP collections [14–17]. High-throughput computational screening is already well established for metal-organic frameworks [18–20], and the same general strategy can be applied to discrete molecular cages. This enables a practical first question to be addressed before affinity, uptake, release, or catalytic performance are considered. Which guests are geometrically able to pass through the pore opening and fit within the internal cavity? Experimental work on a PEGylated cuboctahedral copper-paddlewheel MOP that loads and releases 5-fluorouracil provides a clear precedent for the relevance of this size-compatibility problem [21].

Polyoxometalates (POMs) are a particularly relevant guest class for geometric screening. They are nanoscale metal-oxo clusters with substantial variation in structure, composition, charge, and size, and they have been investigated in catalysis, energy conversion, medicine, and molecular materials [22, 23]. Recent data resources now provide increasingly large collections of POM formulas and structures in forms suitable for computational analysis [24–26]. Experimental POM-in-cage systems have already shown that encapsulation can combine the confined environment of a molecular host with the redox or catalytic properties of the cluster [27]. These studies establish the chemical value of POM encapsulation, but the field still lacks a systematic map showing which POMs are geometrically compatible with which MOP hosts.

Here, we address this gap by screening 1,471 MOPs against 1,936 POMs from POMDB [26] and 292 drug and research compounds retrieved from PubChem [28, 29]. This produced 3.28 million pairwise comparisons. Each guest was described by its outer diameter, which was compared with the host pore-opening diameter to assess entry and with the inner-cavity diameter to assess residence. The model uses simple scalar size descriptors and does not predict binding affinity, uptake, stability, release, or reaction outcome. It also does not account for molecular anisotropy, electrostatics, solvation, counterions, or host flexibility. The screen is intended to remove clearly incompatible pairs and identify geometrically plausible candidates for further study by docking, molecular dynamics, electronic-structure calculations, or experiment.

The analysis addresses two questions at the scale of an entire MOP library. The first is whether pore entry or cavity residence is the main geometric constraint on host–guest compatibility. The second is whether particular MOP topologies fall within the size range needed for close geometric matching with POM guests. These questions have direct practical relevance. If pore openings are the main bottleneck, synthetic design should focus on widening or modifying the windows instead of only increasing cavity size. If certain MOP topologies are strongly enriched in close-fitting POM matches, they provide rational starting points for more detailed computational studies and experimental validation. Although developed here for MOPs, the workflow can also be applied to coordination cages and organic cages when equivalent geometric descriptors are available.

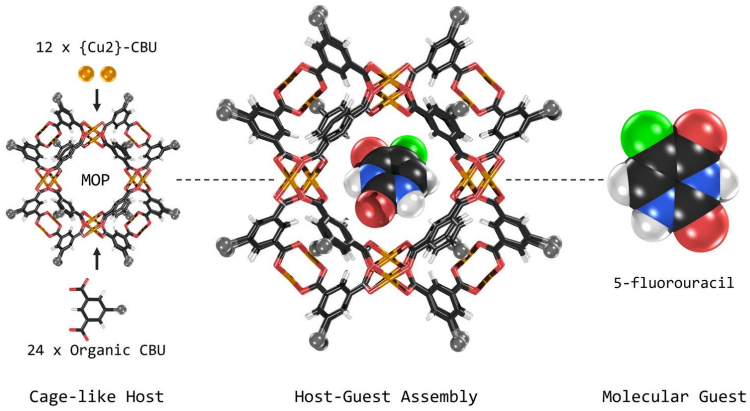


Figure 1: *Assembly of a MOP host from 12 Cu_2 paddlewheel coordination building units and 24 bent organic building units, together with 5-fluorouracil as an experimentally demonstrated host–guest system [21]. Grey spheres indicate the peripheral sites used for PEGylation through click chemistry in the original study.*

2 Methodology

2.1 Data Sources

The MOP dataset comprised 1,471 assembly-generated structures from MOP-DB [30]. The structures were rebuilt using the previously reported automated assembly-modelling workflow [15] and deposited as XYZ files with associated metadata. The metadata include the MOP identifier, assembly model, shape category, formula, molecular weight, outer diameter, and inner-cavity diameter. The inner-cavity diameter was used as deposited.

The maximum adjusted pore size was calculated from the XYZ coordinates and the pore-window centroids provided in `assembly_cycles_and_centroids.json`. For each centroid $\mathbf{c}_k$, the nearest framework atom i was identified and its covalent radius R_i was subtracted from the centroid-to-atom distance. The largest pore-window diameter was then defined as

$$\text{PoreSize} = \max_{k=1,\ldots,N} 2\left(\min_i |\mathbf{c}_k - \mathbf{a}_i| - R_i\right). \tag{1}$$

This scalar descriptor does not capture window anisotropy, pore flexibility, or obstruction by functional groups.

POM structures ($n = 1,936$) were obtained from POM-DB, a computation-ready dataset of experimentally reported polyoxometalate structures [26]. The dataset provides CML structure files and associated metadata, including molecular formula, mass, charge, classification, and InChI identifiers. Geometric descriptors were calculated from the deposited atomic coordinates.

A collection of 292 drug and research compounds was assembled as PubChem compound identifiers and retrieved through the PUG REST API [28, 29]. For each compound, we obtained its molecular structure, SMILES representation, name, and molecular weight. RDKit [31] was used to calculate LogP, topological polar surface area, and the numbers of hydrogen-bond donors, hydrogen-bond acceptors, and rotatable bonds. All 292 PubChem records were successfully retrieved and processed.

2.2 Molecular Descriptor Calculation

The outer diameter (OD) of each POM was defined as the largest distance between any two atoms in the deposited structure

$$\text{OD} = \max_{i,j} |\mathbf{r}_i - \mathbf{r}_j| \tag{2}$$

This descriptor provides a simple measure of the maximum molecular extent [32]. It represents the largest dimension that must be accommodated by a host, although it does not describe molecular anisotropy, aspect ratio, or the minimum cross section required for passage through a pore.

The same definition was used for the drug and research compounds. For each compound, a three dimensional SDF structure was retrieved from PubChem through the PUG REST API, and the maximum interatomic distance was calculated from its conformer coordinates. When no PubChem 3D structure was available, a conformer was generated using RDKit's ETKDGv3 method and optimised with the MMFF94 force field. PubChem conformers were available for 291 of the 292 compounds, while one conformer was generated with RDKit.

The mass weighted radius of gyration (RG) of each POM was calculated as

$$\text{RG} = \sqrt{\frac{\sum_{i=1}^{N} m_i \left|\mathbf{r}_i - \mathbf{r}_{\text{cm}}\right|^2}{\sum_{i=1}^{N} m_i}} \tag{3}$$

where m_i and $\mathbf{r}_i$ are the mass and position of atom i, and $\mathbf{r}_{\text{cm}}$ is the centre of mass. RG was used only to describe molecular size and examine correlations. It correlates strongly

with OD across the POM dataset, with $r = 0.89$ and $R^2 = 0.79$, but it does not enter the compatibility criteria.

RDKit [31] was also used to calculate LogP, topological polar surface area, hydrogen bond donor and acceptor counts, and the number of rotatable bonds from the compound SMILES strings. For descriptive figures, compound RG was approximated as $\mathrm{RG} \approx \mathrm{OD}/2.5$. These values provide a consistent first description of the compound set, although a single conformer does not represent the full conformational range of a flexible molecule.

2.3 Systematic Screening

Each guest was evaluated against every MOP using two geometric criteria. Entry was considered possible when the guest outer diameter did not exceed the largest pore opening diameter of the host

$$\mathrm{OD}_{\mathrm{guest}} \leq \mathrm{PoreSize}_{\mathrm{MOP}} \tag{4}$$

Residence was considered possible when the guest outer diameter did not exceed the inner cavity diameter

$$\mathrm{OD}_{\mathrm{guest}} \leq \mathrm{InnerDiameter}_{\mathrm{MOP}} \tag{5}$$

Each pair was assigned to one of four outcomes according to whether it satisfied both criteria, only the entry criterion, only the residence criterion, or neither criterion. The full screening included $1{,}471 \times 1{,}936 = 2{,}848{,}256$ POM–MOP pairs and $1{,}471 \times 292 = 429{,}532$ compound–MOP pairs, giving 3.28 million comparisons in total.

The two criteria describe different geometric requirements, although their influence is not equal in this dataset. For 1,458 of the 1,471 MOPs, the pore opening is smaller than the inner cavity. A guest that satisfies the entry criterion will therefore usually also satisfy the residence criterion. The pore opening consequently acts as the main geometric constraint for most hosts in the dataset. Pairs that satisfy only the residence criterion remain relevant because they may represent guests introduced during cage assembly instead of through an existing pore.

2.4 Close-Fitting Pair Selection

For every compatible POM–MOP pair, cavity filling was calculated from the guest outer diameter and the host inner cavity diameter

$$\text{Cavity filling } (\%) = \frac{\mathrm{OD}_{\mathrm{guest}}}{\mathrm{InnerDiameter}_{\mathrm{MOP}}} \times 100 \tag{6}$$

Pairs with cavity filling between 70 and 90% were retained as close-fitting candidates. This range was used to identify guests that occupy a large fraction of the cavity while

retaining some geometric clearance. It is not an experimentally established optimum for binding or encapsulation.

2.5 Statistical Analysis

Compatibility was summarised from both the guest and host perspectives. For each guest, we calculated the number and fraction of MOPs satisfying the entry and residence criteria. For each MOP, we calculated the number and fraction of compatible POMs and compounds. Size distributions and relationships between molecular descriptors were examined using histograms, cumulative distributions, density plots, and Pearson correlation coefficients.

The distribution of close-fitting POM candidates across MOP assembly models was evaluated using one-sided Mann–Whitney U tests, comparing the number of close-fitting pairs per host between the selected topology or shape group and the remaining MOPs in the dataset. This quantity is strongly zero-inflated, since 1,420 of the 1,471 hosts form no close-fitting pairs at all. Under this degree of tying the normal approximation to the null distribution of the U statistic is unreliable and yields anticonservative tail probabilities, so the exact permutation distribution was used for the assembly-model comparison, where the smaller group contains 20 hosts. For the shape-category comparison the exact distribution is not computationally accessible, and a Monte Carlo permutation test with 10^5 random reassignments of the group labels was used instead. Effect sizes are reported as the rank-biserial correlation, which equals $2U/(n_1 n_2) - 1$. As an independent robustness check, a one-sided Fisher exact test was applied to the binary outcome of whether a host forms any close-fitting pair. All tests were performed with SciPy [33]. Reported percentages were calculated from the corresponding pair or structure counts. Figures were generated in Python with NumPy [34] and Matplotlib [35] and exported as vector PDF files and 300 DPI PNG files.

3 Results

3.1 Dataset Overview

Figure 2 compares the size distributions of the three datasets on a common diameter scale. The figure includes MOP inner-cavity diameters, MOP pore-opening diameters, POM outer diameters, and the outer diameters of the drug and research compounds. Each distribution was independently normalised to its maximum so that the shapes of the curves can be compared without being affected by dataset size.

The overlap between MOP cavity sizes and the smaller POMs defines the size range in which encapsulation is geometrically possible. For 99.1% of the MOPs, the pore opening is smaller than the inner cavity. Pore entry therefore acts as the main geometric constraint in this dataset. The 1,471 MOPs have outer diameters between 14.6 and 76.6 Å, with a median of 34.7 Å. Their inner-cavity diameters reach 42.95 Å. The 1,936 POM structures cover a broad size range, with the highest density near 15 Å and a smaller number of

much larger structures. Their mass-weighted radii of gyration correlate strongly with outer diameter, with $r = 0.89$ and $R^2 = 0.79$. The 292 drug and research compounds occupy the small-to-intermediate size range and have a median outer diameter of 15.1 Å.

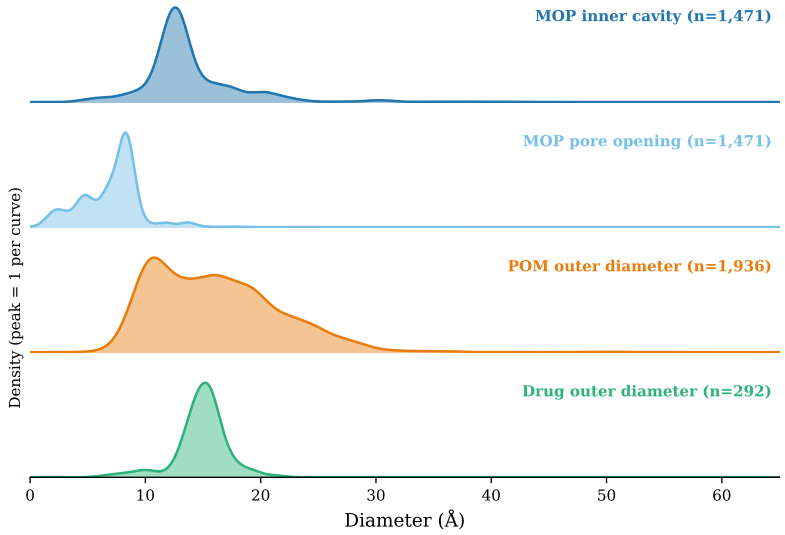


Figure 2: *Joint dataset overview: size distributions of MOP inner cavity diameters and pore openings, POM outer diameters, and 3D-conformer-derived drug outer diameters, stacked on a shared diameter axis (each row independently peak-normalised, so height compares curve shape, not sample size; n per series is given).*

3.2 Assembly Model Distribution

The 1,471 MOPs belong to 18 assembly models, labelled AM_1 to AM_18, and nine broader shape categories [14]. AM_16 and AM_17 each contain 400 MOPs and together account for more than half of the dataset. AM_11 is the next largest group with 210 MOPs. The largest median inner cavity diameters occur in AM_15, AM_18, and AM_13, with values of 31.4, 29.8, and 26.7 Å, respectively. AM_13 contains only five MOPs and shares the Octahedron shape category with several assembly models that have much smaller cavities. This shows that the broad shape category does not fully describe cavity size or guest compatibility. Table 1 lists the number of MOPs in each assembly model and the number of POMs and compounds that are compatible with at least one MOP in that group. AM_13 and AM_5 show the highest compatibility counts despite containing only five and 20 MOPs. AM_7 and AM_12 show no compatible POMs or compounds.

3.3 Compound Dataset Characterisation

The 292 drug and research compounds are predominantly drug-like. Of these, 207 compounds (70.9%) have no Lipinski violations, 82 compounds (28.1%) have one, and 3

Table 1: *Number of MOPs and compatible guests for each assembly model. The POM and compound columns report how many of the 1,936 POMs and 292 compounds are compatible with at least one MOP in the corresponding group.*

Assembly Model	# MOPs	# POMs	# Compounds
AM_1	18	5	3
AM_2	30	111	9
AM_3	11	2	1
AM_4	84	663	31
AM_5	20	1,255	263
AM_6	87	81	9
AM_7	18	0	0
AM_8	34	3	2
AM_9	90	794	72
AM_10	15	6	4
AM_11	210	289	16
AM_12	5	0	0
AM_13	5	1,895	280
AM_14	6	33	6
AM_15	10	114	9
AM_16	400	752	54
AM_17	400	732	48
AM_18	28	129	11

compounds (1.0%) have two. The few violations arise mainly from molecular weights above 500 Da, as shown in Figure 3c. The compound set therefore covers a broad small-molecule size range, and the 3D-conformer-derived OD values are not dominated by outliers.

3.4 POM Screening Results

Of the 1,936 POMs, 1,895 (97.9%) can enter at least one MOP and also fit within its cavity. A total of 1,931 POMs (99.7%) can fit within the cavity of at least one MOP. On average, each POM can enter 40.1 MOPs with a standard deviation of 86.7. The average number of MOP cavities that can accommodate a POM is 562.5 with a standard deviation of 523.5. The mean number of hosts a POM can enter is identical to the mean number for which it satisfies both criteria, which confirms that pore opening rather than cavity size is the limiting constraint. The smallest POMs are compatible with the largest number of hosts. POM_X5358 has an outer diameter of 5.59 Å and is compatible with 1,053 MOPs, which represents 71.6% of the dataset. Compatibility decreases rapidly as POM size increases. POMs larger than 25 Å are compatible with no more than one MOP. Only 11 of the 52 POMs larger than 28 Å have a compatible host. The largest POM satisfying both criteria has an outer diameter of 28.14 Å.

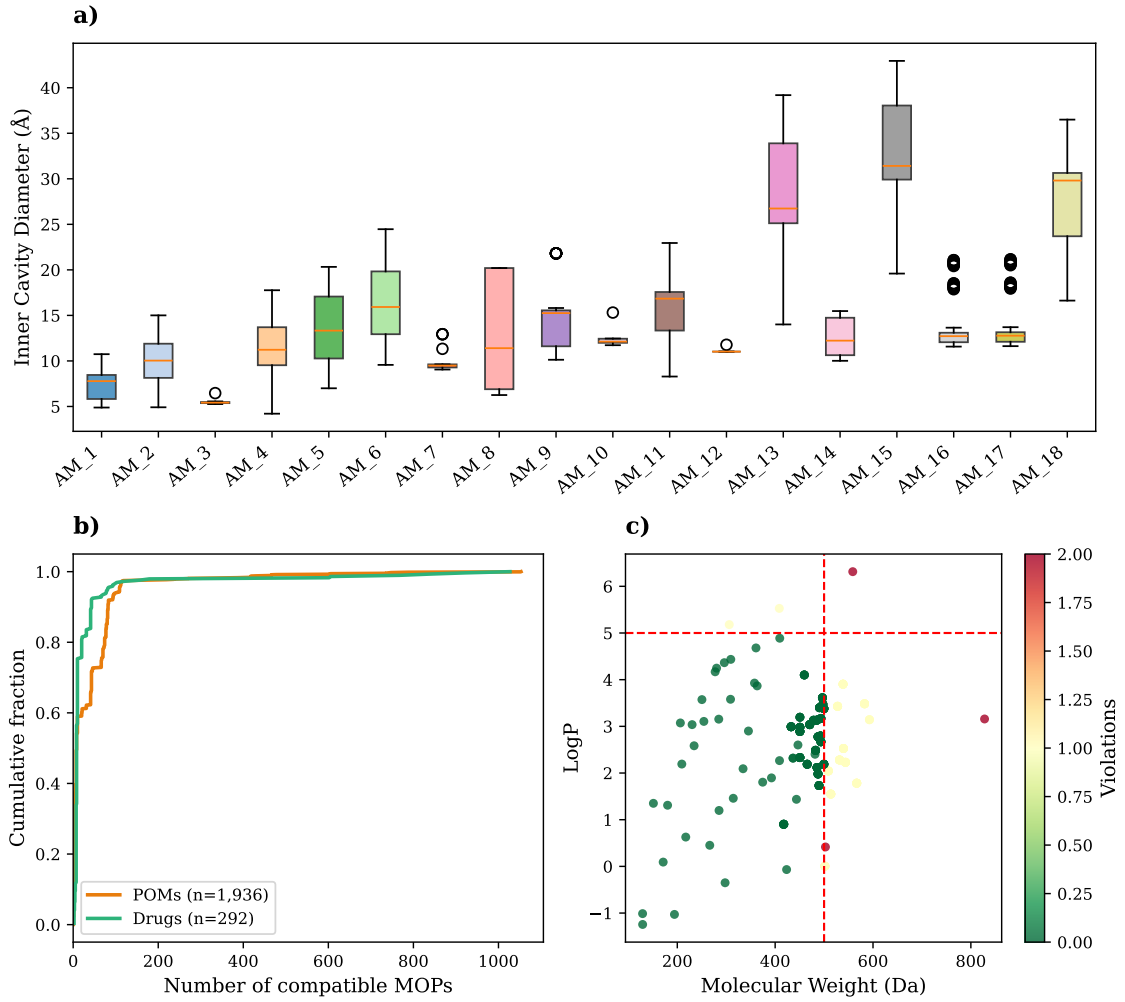


Figure 3: *MOP structural diversity and guest compatibility. (**a**) Inner cavity diameter distributions for assembly models AM_1 to AM_18 following [14]. AM_15 and AM_18, both classified as Icosahedron, and AM_13, classified as Octahedron, exhibit the largest cavities. AM_13 shares its shape label with several assembly models that have much smaller cavities, including AM_5, AM_6, AM_8, and AM_11. The number of MOPs in each assembly model is given in Table 1. (**b**) Cumulative compatibility distributions for POM guests and compounds. Most compounds reach high cumulative fractions at low compatible-MOP counts, while the POM population has a longer tail driven by the smallest clusters. (**c**) Compound molecular weight versus LogP coloured by the number of Lipinski Rule of Five violations. Dashed lines mark molecular weight of 500 Da and LogP of 5. Violation counts are given in the text.*

Table 2: *POMs with the highest MOP compatibility.*

Guest ID	Formula	OD (Å)	# MOPs	% of Dataset
POM_X5358	$[V_3O_9]^{3-}$	5.59	1,053	71.6%
POM_X5357	$[V_3O_9]^{3-}$	5.62	1,052	71.5%
POM_X5351	$[V_3O_7]^{-}$	6.51	972	66.1%
POM_X5147	$[V_4O_{12}]^{4-}$	7.41	788	53.6%
POM_X4972	$[Mo_5O_{17}]^{4-}$	7.49	773	52.5%

3.5 MOP Host Results

The MOPs with the largest number of compatible guests generally have the largest pore openings and cavities. AM_13 shows the highest overall compatibility. This assembly model has the form (3-planar)$_8$(2-bent)$_{12}$ with O_h symmetry and inner cavity diameters between 25 and 39 Å. A different pattern appears when the hosts are ranked by the number of close fitting POMs with cavity filling between 70 and 90%. The hosts with the highest ranks belong to AM_5, which has the form (3-planar)$_4$(2-bent)$_6$ with T_d symmetry. AM_5_21 forms 444 close fitting pairs. AM_5_23 and AM_5_24 each form 405 pairs. AM_5_9 forms 390 pairs. These hosts accommodate fewer POMs overall than AM_13, but their cavity dimensions more closely match the main size range of the POM dataset. All eight hosts with the highest ranks belong to AM_5.

Table 3: *Top MOP hosts ranked by the number of close fitting POM pairs with 70–90% cavity filling. All listed hosts belong to AM_5, which corresponds to the (3-planar)$_4$(2-bent)$_6$ assembly model with T_d symmetry. The assembly model column is therefore omitted.*

MOP_ID	Formula	OD (Å)	Inner Diam. (Å)	# Compat. Guests	Close-Fit. Pairs
AM_5_21	$[C_{324}\ H_{252}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	49.05	20.33	1,255	444
AM_5_23	$[C_{360}\ H_{288}\ Mo_{12}\ N_{24}\ O_{48}]^{12-}$	49.61	19.48	1,170	405
AM_5_24	$[C_{360}\ H_{288}\ N_{24}\ O_{48}\ Rh_{12}]^{12-}$	49.61	19.48	1,170	405
AM_5_9	$[C_{252}\ H_{204}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	38.44	14.15	548	390
AM_5_17	$[C_{276}\ H_{204}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	44.53	17.70	980	379
AM_5_19	$[C_{312}\ H_{240}\ Mo_{12}\ N_{24}\ O_{48}]^{12-}$	45.26	16.86	887	365
AM_5_20	$[C_{312}\ H_{240}\ N_{24}\ O_{48}\ Rh_{12}]^{12-}$	45.26	16.86	887	365
AM_5_5	$[C_{240}\ H_{204}\ Cu_{12}\ N_{36}\ O_{48}]^{12+}$	37.07	13.35	473	354

3.6 Compatibility Analysis

The number of compatible MOPs per guest is strongly skewed toward low values. Most guests are compatible with only a small number of MOPs, while the smallest guests are compatible with more than 1,000 hosts. All 292 compounds are compatible with at least one MOP at the molecule level. The pair-level picture is much more selective. Only 9,564 of 429,532 compound–MOP combinations, or 2.23%, satisfy the entry criterion. This reflects the size distribution of the host library, because most of the 1,471 MOPs have

pore openings below the 3D-conformer-derived outer diameter range of the compounds. The high molecule-level compatibility therefore comes from a small number of large-pore hosts, not from a broad size match across the full library. The cumulative distributions in Figure 3b show this difference clearly. Most compounds reach high cumulative fractions at low compatible-MOP counts, while the POM distribution has a longer tail driven by the smallest clusters.

Figure 4 shows the same trends in pair-space. In panel (a), colour represents the relative density of compatible POM–MOP pairs across POM outer-diameter and MOP pore-size bins. Empty cells are shown in white, and the dashed diagonal marks the OD = pore-size boundary. Of the 77,692 compatible POM–MOP pairs, 5,688 or 7.3% fall within the 70–90% cavity-filling range defined in Section 3.7. Panel (b) shows the corresponding map for compounds. Compatibility is close to zero for MOP pore openings below about 11 Å. It rises strongly in the 12–18 Å range and reaches 1.0 above about 22 Å, largely independent of compound size. Only the transition cells between fully incompatible and fully compatible regions are annotated.

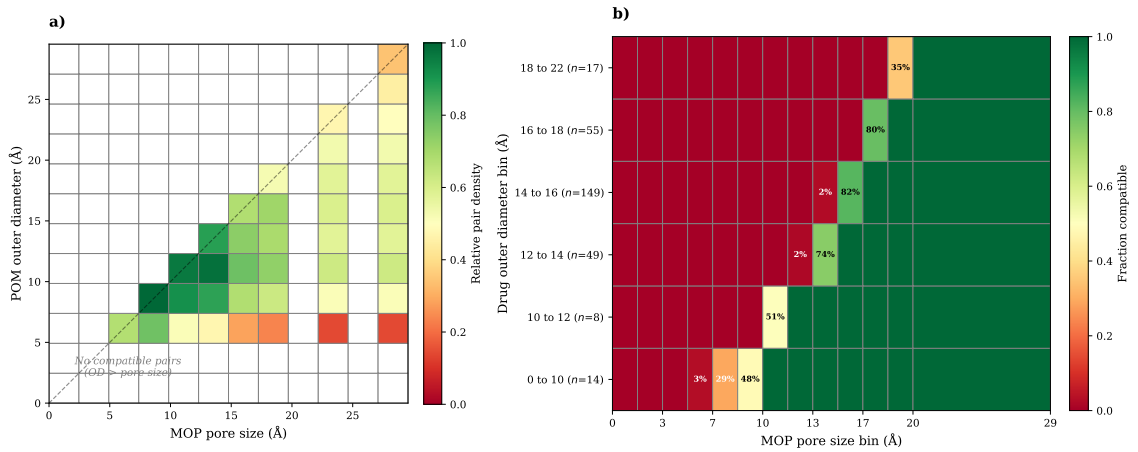


Figure 4: *Compatibility maps for POM–MOP and compound–MOP pairs. (**a**) Relative density of compatible POM–MOP pairs across POM outer-diameter and MOP pore-size bins. White cells contain no compatible pairs. The dashed diagonal marks OD = pore size. (**b**) Fraction of compatible compound–MOP combinations across compound outer-diameter and MOP pore-size bins. Only transition cells between 0% and 100% compatibility are annotated. The y-axis labels give the number of compounds in each outer-diameter bin.*

3.7 Close-Fitting Pairs

We identified 5,688 POM–MOP pairs, corresponding to 7.3% of all compatible pairs, in which the guest outer diameter fills 70–90% of the MOP inner cavity. These pairs define a practical priority subset with a closer geometric match than most compatible pairs, although this range is not intended as a proven optimum for binding. Table 4 presents representative examples across the full 70–90% filling interval and across several MOP

families. These examples show that close geometric matching is possible for POMs ranging from small clusters of about 5 Å to medium-sized polyoxometalates of about 18 Å.

Table 4: *Representative close-fitting POM–MOP pairs spanning the 70–90% cavity filling range. The pairs are (guest, MOP): (POM_X5357, AM_1_11), (POM_X493, AM_5_17), (POM_X3736, AM_5_20), (POM_X1754, AM_5_21), (POM_X4384, AM_5_23), (POM_X509, AM_5_9), (POM_X277, AM_5_12), and (POM_X5220, AM_4_66).*

Guest Formula	MOP Formula	Guest OD (Å)	MOP Inner Diam. (Å)	Fill (%)
$[V_3\ O_9]^{3-}$	$[C_{60}\ H_{78}\ O_{56}\ S_5\ V_{12}]^{10-}$	5.62	6.29	89.4
$[Cu_2\ Mo_8\ O_{26}\ C_{12}\ N_8]^{2-}$	$[C_{276}\ H_{204}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	15.64	17.70	88.4
$[Re_2\ W_8\ O_{36}\ C_6]^{10-}$	$[C_{312}\ H_{240}\ N_{24}\ O_{48}\ Rh_{12}]^{12-}$	14.85	16.86	88.1
$[V_6\ O_{25}\ C_{24}\ N_2]^{2-}$	$[C_{324}\ H_{252}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	17.79	20.33	87.5
$[Cr\ Mn_4\ W_{18}\ O_{70}\ Si_2]^{7-}$	$[C_{360}\ H_{288}\ Mo_{12}\ N_{24}\ O_{48}]^{12-}$	17.01	19.48	87.3
$[Mo_6\ O_{18}\ C_{12}\ N]^{2-}$	$[C_{252}\ H_{204}\ Cu_{12}\ N_{24}\ O_{48}]^{12+}$	12.05	14.15	85.1
$[V_6\ O_{19}\ C_{10}]^{2-}$	$[C_{288}\ H_{240}\ N_{24}\ O_{48}\ Rh_{12}]^{12-}$	11.28	13.31	84.7
$[W_{11}\ O_{39}\ Si]^{8-}$	$[C_{108}\ H_{132}\ N_{24}\ O_{76}\ V_{28}]^{0}$	10.42	13.83	75.4

4 Discussion

4.1 Relative Importance of Entry and Cavity Size

In 99.1% of the MOPs, the pore opening is smaller than the inner cavity. Pore entry therefore acts as the effective size limit for almost the entire library.

This result has a direct design implication. If the aim is to expand encapsulation compatibility, synthetic effort should focus first on pore engineering. Enlarging the cavity alone will not help if guests still cannot pass through the window. Larger cages can increase the number of compatible guests, though pore passage remains the main geometric barrier. A different balance between the two criteria may appear in future datasets that contain MOPs with unusually large windows relative to their cavities. For the present library, however, pore opening is the dominant control variable.

The model remains intentionally simple. Outer diameter is a single scalar descriptor based on the two atoms that are farthest apart. It therefore cannot distinguish between a compact guest and an elongated or flattened one if they share the same maximum extent. The pore opening is also treated as a fixed circular size, even though real windows may be elliptical, irregular, or partly blocked by functional groups or counterions. The entry and residence criteria are applied at exact equality, so the model does not account for pore flexibility, guest orientation, or solvent effects. These simplifications are what make a screen of 3.28 million pairs feasible, and they also define its scope. The present filter identifies pairs that are geometrically plausible. It does not predict which pairs will actually bind.

4.2 Assembly Model Concentration of Close-Fitting Pairs

The 5,688 close-fitting POM–MOP pairs are distributed very unevenly across the MOP library. A total of 4,459 pairs, corresponding to 78%, come from only 14 active hosts, and all of these hosts belong to the same assembly model, AM_5. This topology has the form (3-planar)$_4$(2-bent)$_6$ with T_d symmetry and is classified within the broader Octahedron shape category.

The broader shape label does not capture this pattern well. In this dataset, the Octahedron category contains five different assembly-model topologies and 356 MOPs in total. These topologies are also very unevenly represented. One of them, the O_h-symmetric topology (4-pyramidal)$_6$(2-linear)$_{12}$, accounts for 210 MOPs, whereas the T_d-symmetric AM_5 topology accounts for only 20. When all octahedral MOPs are grouped together, AM_5 is mixed with four other topologies that contribute very little to the close-fitting pairs. This weakens the statistical signal. A one-sided Mann–Whitney U test comparing the 356 octahedral MOPs with the remaining 1,115 gives $p = 0.087$ by Monte Carlo permutation, which is not significant at conventional levels. The same test applied to the correct grouping variable, namely the 20 AM_5 MOPs against the other 1,451, gives $U = 24{,}503$ with an exact one-sided $p = 4.1 \times 10^{-9}$ and a rank-biserial correlation of $r = 0.69$. The same contrast expressed as a binary outcome gives the same picture, since 14 of the 20 AM_5 hosts form at least one close-fitting pair compared with 37 of the other 1,451, corresponding to a Fisher exact odds ratio of 89 and $p = 1.8 \times 10^{-17}$. The six inactive AM_5 hosts are those whose cavities fall outside the POM size range, so the enrichment reflects a subset of the topology and not every member of it. The enrichment therefore becomes clear only when the analysis is carried out at the assembly-model level. For this dataset, assembly-model identity is more informative than the coarse shape category.

A second topology also contributes substantially. AM_4, which has the form (3-pyramidal)$_4$(2-linear)$_6$ with T_d symmetry and is classified as Tetrahedron, contributes 992 pairs, or 17%, from 15 active hosts. By contrast, AM_13, which is also labelled Octahedron and has the form (3-planar)$_8$(2-bent)$_{12}$ with O_h symmetry, contributes only 38 pairs, or 0.7%, from two hosts. Together, AM_5 and AM_4 account for 96% of all close-fitting pairs while representing only 29 of the 1,471 MOPs in the library, corresponding to 2.0%. The active AM_5 hosts have inner cavity diameters between 12.5 and 20.3 Å, which closely overlap with the peak of the POM outer-diameter distribution shown in Figure 2. This size overlap is the simplest explanation for the strong concentration of close-fitting pairs in AM_5 and AM_4. It does not imply that octahedral or tetrahedral geometries are inherently superior in general. It shows that these particular assembly models occupy the size range that best matches the bulk of the POM dataset. Two further points should be kept in mind when interpreting this result. First, AM_5_23 and AM_5_24 have identical inner cavity diameters and identical compatibility counts, as shown in Table 3, which indicates the presence of near-duplicate hosts in the library. Second, 68.7% of all 1,471 MOPs belong to only three assembly models, as shown in Table 1. The library is therefore structurally concentrated and is not a uniform sample of MOP space. A deeper topology-level explanation would require richer shape descriptors and scaffold-level deduplication, which are beyond the scope of the present study. The main result here is that large-scale integration of public datasets makes it possible to identify this concentration clearly and to show that assembly-model identity is the correct level for topology-based enrichment analysis.

4.3 Interpreting Compound Compatibility and Retrospective Validation

All 292 drug and research compounds satisfy the geometric filter for at least one MOP. Taken on its own, this could suggest broad applicability for molecular encapsulation. The pair-level result is much more selective. Only 9,564 of 429,532 compound–MOP combinations satisfy the entry criterion, corresponding to 2.2%. This reflects the size distribution of the MOP library. A small number of hosts with large pore openings can accommodate many compounds, while most MOPs have pore openings below the compound size range. The result is therefore not broad compatibility across the full host set. It is selective compatibility concentrated in a limited number of larger-pore MOPs. This is still useful as a first filter because it narrows the experimental search space substantially. The 3D conformer based outer diameters should nevertheless be interpreted as geometric plausibility estimates, since a single conformer does not represent the full conformational range of a flexible molecule. The strongest validation of this geometric filter would be a retrospective benchmark against a curated dataset of experimentally confirmed MOP host–guest systems. No such dataset is currently available at the scale of the present screen. Two literature examples are nevertheless consistent with the present framework. The cuboctahedral Cu paddlewheel MOP reported by Zhao *et al.*[21] was shown to load 5-fluorouracil, and this is geometrically consistent with a small guest inside a host with sufficiently large pore openings. The exact MOP used in that study is not present in the current library, so the pair cannot be counted as a direct validation case. The POM-in-cage system reported by Cai *et al.*[27] likewise shows that encapsulation of polyoxometalates in synthetic cages is experimentally feasible, although the specific cage and POM identifiers differ from those used here. A rigorous retrospective benchmark therefore remains future work and will require community curation of experimentally verified host–guest pairs.

5 Conclusions

Geometric screening of 3.28 million host–guest combinations across 1,471 MOPs, 1,936 polyoxometalates, and 292 drug and research compounds leads to two main conclusions. First, pore opening acts as the effective size limit for 99.1% of the MOPs in this library. Within the present scalar model, this means that host accessibility is controlled mainly by the window size and not by the cavity size alone. This points to pore engineering as the most direct route for expanding host compatibility. Second, close fitting POM–MOP pairs are concentrated very strongly in one assembly model. The topology (3-planar)$_4$(2-bent)$_6$ with T_d symmetry contributes 78% of the 5,688 pairs with 70–90% cavity filling, even though only 14 hosts are involved ($p = 4.1 \times 10^{-9}$, rank-biserial $r = 0.69$). A second topology contributes a further 17%. These hosts have inner cavity diameters between 12.5 and 20.3 Å, which closely match the main size range of the POM dataset. The results provide an open and computationally tractable first screening layer for prioritising MOP host–guest candidates.

The next step is retrospective validation against a curated set of experimentally confirmed

encapsulation systems. That would turn the present geometric filter into a benchmarked screening stage. Methodologically, the most useful extension would be to move beyond a single outer-diameter descriptor and include multi-axis shape descriptors, pore-window shape, and electrostatic complementarity. Even in its present form, the workflow is easy to extend because it requires only pore size, inner cavity diameter, and guest size. This makes it applicable not only to MOPs but also to coordination cages and organic cages. The interactive web application shown in Figure 5 allows direct querying of the dataset by molecule name or formula and is available at https://digital-chemistry.io/pom-mop/. Source code is available at https://github.com/digital-chemistry/POM-MOP. The present model is intentionally simple. It identifies geometrically plausible candidates, not experimentally validated host–guest systems, and its main value is to reduce a very large search space to a focused set of targets for more detailed computational and experimental study.

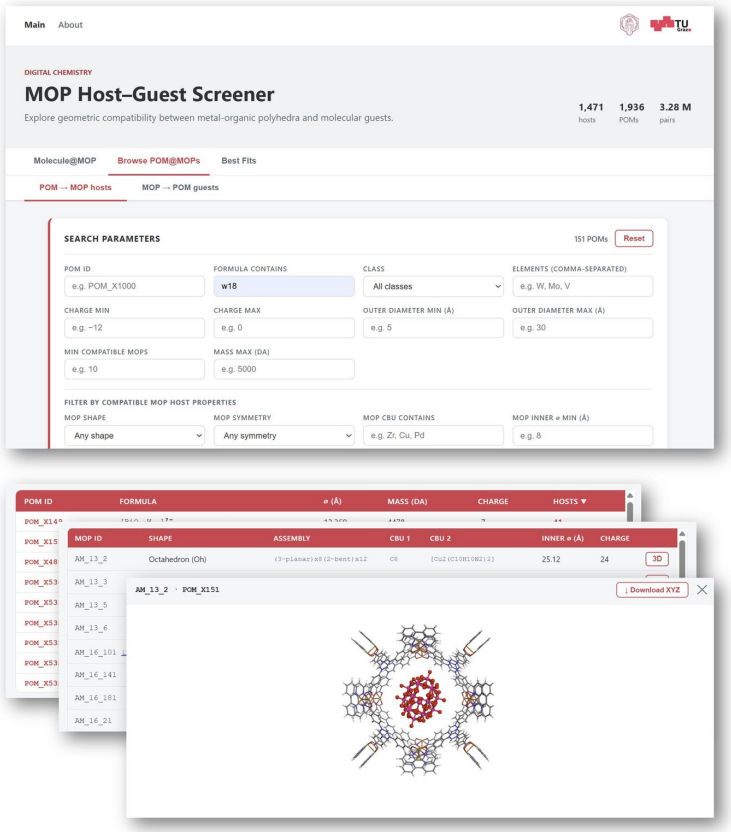


Figure 5: *Interactive web application for host–guest compatibility screening. The figure shows a POM browsing workflow in which filtering by formula (`W18`) returns matching POMs and their compatible MOP hosts, and selecting a pair opens a 3D view of the corresponding host–guest assembly (AM_13_2 with POM_X151). The guest is shown in ball-and-stick representation and the host in stick representation. Source code is available at https://github.com/digital-chemistry/POM-MOP.*

Author Contributions

Conceptualisation, A.K. and D.T. Software, A.K., L.A., G.Bi. and G.Bo. Data curation, A.K. Writing of the original draft, A.K. Review and editing, all authors. Funding acquisition, A.K. and D.T. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

A.K. thanks the Field of Expertise Advanced Materials Science of Graz University of Technology for Competitive Start-up Funding support under project AF22-635-1. D.T. acknowledges the support from the Ministry of Education and Science of the Republic of North Macedonia through the project "Utilizing AI and National Large Language Models to Advance Macedonian Language Capabilities". D.T. also acknowledges cofunding from the EC/EuroHPC JU, the Ministry of Digital Transformation of the Republic of North Macedonia, and the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, through the VEZILKA project, Grant Agreement No. 101263128. Supported by the TU Graz Open Access Publishing Fund.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All code required to reproduce the analysis is available in the accompanying GitHub repository at https://github.com/digital-chemistry/POM-MOP under the MIT License. The repository includes scripts for descriptor calculation, pore-size computation, and systematic screening, the interactive web application, and derived CSV and JSON tables covering MOP and guest descriptors, compound properties, and the full pairwise compatibility matrices for entry, residence, and combined compatibility of POM and compound guests. Raw structure collections are archived separately with persistent DOIs. The 1,471 MOP hosts are available as MOP-DB [30]. POM structures are available from POM-DB [26]. The 292 drug and research compounds are included as a derived table in the repository, including PubChem-derived 3D conformer outer diameters for 291 compounds and one RDKit-generated conformer. PubChem is also queried live through the web application for on-demand molecule-name searches.

Abbreviations

MOP, Metal-Organic Polyhedron; POM, Polyoxometalate; OD, Outer Diameter; RG, Radius of Gyration; TPSA, Topological Polar Surface Area; SMILES, Simplified Molecular Input Line Entry System.

Correspondence

Corresponding authors: Aleksandar Kondinski (kondinski@tugraz.at) and Dimitar Trajanov (dimitar.trajanov@finki.ukim.mk).

References

[1] Yoshizawa, M.; Klosterman, J.K.; Fujita, M. Functional Molecular Flasks: New Properties and Reactions within Discrete, Self-Assembled Hosts. *Angew. Chem. Int. Ed.* **2009**, *48*, 3418–3438. 10.1002/anie.200805340

[2] Ajami, D.; Liu, L.; Rebek Jr., J. Soft Templates in Encapsulation Complexes. *Chem. Soc. Rev.* **2015**, *44*, 490–499. 10.1039/C4CS00065J

[3] Rebek Jr., J. Molecular Behavior in Small Spaces. *Acc. Chem. Res.* **2009**, *42*, 1660–1668. 10.1021/ar9001203

[4] Pluth, M.D.; Bergman, R.G.; Raymond, K.N. Proton-Mediated Chemistry and Catalysis in a Self-Assembled Supramolecular Host. *Acc. Chem. Res.* **2009**, *42*, 1650–1659. 10.1021/ar900118t

[5] Bruns, C.J.; Fujita, D.; Hoshino, M.; Sato, S.; Stoddart, J.F.; Fujita, M. Emergent Ion-Gated Binding of Cationic Host–Guest Complexes within Cationic $M_{12}L_{24}$ Molecular Flasks. *J. Am. Chem. Soc.* **2014**, *136*, 12027–12034. 10.1021/ja505296e

[6] Han, M.; Engelhard, D.M.; Clever, G.H. Self-Assembled Coordination Cages Based on Banana-Shaped Ligands. *Chem. Soc. Rev.* **2014**, *43*, 1848–1860. 10.1039/C3CS60473J

[7] Percástegui, E.G.; Ronson, T.K.; Nitschke, J.R. Design and Applications of Water-Soluble Coordination Cages. *Chem. Rev.* **2020**, *120*, 13480–13544. 10.1021/acs.chemrev.0c00672

[8] McConnell, A.J. Metallosupramolecular Cages: From Design Principles and Characterisation Techniques to Applications. *Chem. Soc. Rev.* **2022**, *51*, 2957–2971. 10.1039/D1CS01143J

[9] Montà-González, G.; Sancenón, F.; Martínez-Máñez, R.; Martí-Centelles, V. Purely Covalent Molecular Cages and Containers for Guest Encapsulation. *Chem. Rev.* **2022**, *122*, 13636–13708. 10.1021/acs.chemrev.2c00198

[10] Yang, X.; Ullah, Z.; Stoddart, J.F.; Yavuz, C.T. Porous Organic Cages. *Chem. Rev.* **2023**, *123*, 4602–4634. 10.1021/acs.chemrev.2c00667

[11] Lee, S.; Jeong, H.; Nam, D.; Lah, M.S.; Choe, W. The Rise of Metal-Organic Polyhedra. *Chem. Soc. Rev.* **2021**, *50*, 528–555. 10.1039/D0CS00443J

[12] Khobotov-Bakishev, A.; Hernández-López, L.; von Baeckmann, C.; Albalad, J.; Carné-Sánchez, A.; Maspoch, D. Metal-Organic Polyhedra as Building Blocks for Porous Extended Networks. *Adv. Sci.* **2022**, *9*, 2104753. 10.1002/advs.202104753

[13] Albalad, J.; Hernández-López, L.; Carné-Sánchez, A.; Maspoch, D. Surface Chemistry of Metal-Organic Polyhedra. *Chem. Commun.* **2022**, *58*, 2443–2454. 10.1039/D1CC07034G

[14] Kondinski, A.; Menon, A.; Nurkowski, D.; Farazi, F.; Mosbach, S.; Akroyd, J.; Kraft, M. Automated Rational Design of Metal-Organic Polyhedra. *J. Am. Chem. Soc.* **2022**, *144*, 11713–11728. 10.1021/jacs.2c03402

[15] Kondinski, A.; Oyarzún-Aravena, A.M.; Rihm, S.D.; Bai, J.; Mosbach, S.; Akroyd, J.; Kraft, M. Automated Assembly Modeling of Metal-Organic Polyhedra. *Eur. J. Inorg. Chem.* **2025**, *28*, e202500115. 10.1002/ejic.202500115

[16] Butler, P.W.V.; Subramanian, A.; Rihm, S.D.; Fischer, A.F.; Choksi, T.S.; Kondinski, A.; Mosbach, S.; Akroyd, J.; Kraft, M. Computational Modelling Workflows for Metal–Organic Polyhedra in The World Avatar. *Phys. Chem. Chem. Phys.* **2026**, *28*, 14828–14838. 10.1039/D5CP04871K

[17] Butler, P.W.V.; Rihm, S.D.; Mosbach, S.; Akroyd, J.; Kraft, M. Inverse Design of Metal-Organic Polyhedra through Molecular Fragmentation and Evolutionary Optimisation. *J. Chem. Inf. Model.* **2026**, *66*, 3933–3943. 10.1021/acs.jcim.5c02956

[18] Colón, Y.J.; Snurr, R.Q. High-Throughput Computational Screening of Metal-Organic Frameworks. *Chem. Soc. Rev.* **2014**, *43*, 5735–5749. 10.1039/C4CS00070F

[19] Chung, Y.G.; Camp, J.; Haranczyk, M.; Sikora, B.J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O.K.; Sholl, D.S.; Snurr, R.Q. Computation-Ready, Experimental Metal-Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185–6192. 10.1021/cm502594j

[20] Daglar, H.; Keskin, S. Recent Advances, Opportunities, and Challenges in High-Throughput Computational Screening of MOFs for Gas Separations. *Coord. Chem. Rev.* **2020**, *422*, 213470. 10.1016/j.ccr.2020.213470

[21] Zhao, D.; Tan, S.; Yuan, D.; Lu, W.; Rezenom, Y.H.; Jiang, H.; Wang, L.-Q.; Zhou, H.-C. Surface Functionalization of Porous Coordination Nanocages via Click Chemistry and Their Application in Drug Delivery. *Adv. Mater.* **2011**, *23*, 90–93. 10.1002/adma.201003012

[22] Long, D.-L.; Burkholder, E.; Cronin, L. Polyoxometalate Clusters, Nanostructures and Materials: From Self-Assembly to Designer Materials and Devices. *Chem. Soc. Rev.* **2007**, *36*, 105–121. 10.1039/B502666K

[23] Miras, H.N.; Yan, J.; Long, D.-L.; Cronin, L. Engineering Polyoxometalates with Emergent Properties. *Chem. Soc. Rev.* **2012**, *41*, 7403–7430. 10.1039/C2CS35190K

[24] Kondinski, A.; Gumerova, N.I.; Rompel, A. Curated Polyoxometalate Formula Dataset. *Data* **2024**, *9*, 124. 10.3390/data9110124

[25] Kondinski, A.; Gumerova, N.I.; Rompel, A.; Falcaro, P.; Schreck, T. Data-Driven Polyoxometalate Chemistry. *Chem. Eur. J.* **2025**, *31*, e202501528. 10.1002/chem.202501528

[26] Kondinski, A. Computation-Ready Experimental Polyoxometalate Structural Dataset (Version 1.0.0) [Dataset]. Zenodo **2026**. 10.5281/zenodo.20370381

[27] Cai, L.-X.; Li, S.-C.; Yan, D.-N.; Zhou, L.-P.; Guo, F.; Sun, Q.-F. Water-Soluble Redox-Active Cage Hosting Polyoxometalates for Selective Desulfurization Catalysis. *J. Am. Chem. Soc.* **2018**, *140*, 4869–4876. 10.1021/jacs.8b00394

[28] Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. 10.1093/nar/gkv951

[29] 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*, D1373–D1380. 10.1093/nar/gkac956

[30] Kondinski, A. MOP-DB: Assembly-Generated Metal–Organic Polyhedra [Dataset]. Zenodo **2026**. 10.5281/zenodo.21109457

[31] Landrum, G. RDKit: Open-Source Cheminformatics Software; Version 2025.03.3. Available online: https://www.rdkit.org/ (accessed on 17 June 2026).

[32] Taylor, R.; Kennard, O. The Estimation of Average Molecular Dimensions from Crystallographic Data. *Acta Crystallogr. Sect. B* **1983**, *39*, 517–525. 10.1107/S0108768183002797

[33] Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S.J.; Brett, M.; Wilson, J.; Millman, K.J.; Mayorov, N.; Nelson, A.R.J.; Jones, E.; Kern, R.; Larson, E.; Carey, C.J.; Polat, İ.; Feng, Y.; Moore, E.W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E.A.; Harris, C.R.; Archibald, A.M.; Ribeiro, A.H.; Pedregosa, F.; van Mulbregt, P. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. 10.1038/s41592-019-0686-2

[34] Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M.H.; Brett, M.; Haldane, A.; del Río, J.F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T.E. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. 10.1038/s41586-020-2649-2

[35] Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. 10.1109/mcse.2007.55