MoleQCage: Computational high-throughput screening for molecular caging prediction

Alexander Kravberg,[†,§] Didier Devaurs,[*,‡,§] Anastasiia Varava,[†] Lydia E. Kavraki,[¶] and Danica Kragic[*,†]

[†]*School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, 10044, Sweden*

[‡]*Department of Computer and Information Sciences, University of Strathclyde, Glasgow, G1 1XH, United Kingdom*

[¶]*Department of Computer Science, Rice University, Houston, TX 77005, United States*

[§]*A.K. and D.D. contributed equally to this paper*

E-mail: didier.devaurs@strath.ac.uk; dani@kth.se

Abstract

Although being able to determine whether a host molecule can enclose a guest molecule and form a caging complex could benefit numerous chemical and medical applications, the experimental discovery of molecular caging complexes cannot yet be achieved at scale. Therefore, we propose MoleQCage, a simple tool for the high-throughput screening of host and guest candidates, based on an efficient robotics-inspired algorithm for molecular caging prediction, providing theoretical guarantees and robustness assessment. MoleQCage is distributed as Linux-based software with a graphical user interface, and is available online at `https://hub.docker.com/r/dantrigne/moleqcage` in the form of a Docker container. Documentation and examples are available as Supporting Information and online at `https://hub.docker.com/r/dantrigne/moleqcage`.

Introduction

A molecular caging complex is defined as a pair of molecules in which a so-called host (or cage) features an internal cavity that can enclose a so-called guest, preventing its escape (Figure 1). In this kind of supra-molecular interaction, we can say that the host cages the guest or, dually, that the guest is caged by the host. In synthetic chemistry, a host molecule is usually created with dynamic covalent bonds allowing its self-assembly around a guest molecule and its later disassembly in response to a specific stimulus (such as temperature, pH, or light). This paradigm has produced exciting biochemical applications, for example in targeted drug delivery, virus trapping, or medical imaging.[1–3] Despite its promises, the use of molecular caging complexes remains challenging, with the discovery or synthesis of host molecules being the main bottleneck.[4]

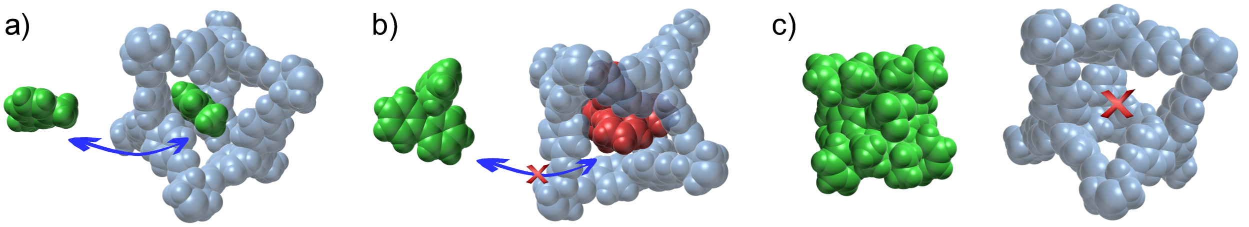

Figure 1: Definition of a molecular caging complex. (a) The guest molecule (in green) can move in and out of the cavity of the host molecule (in blue), and is therefore not caged. (b) The guest fits in the cavity of the host and is either outside (in green) or inside (in red) without the possibility to escape; in this case, the host and guest form a molecular caging complex. (c) The guest cannot fit in the cavity and thus cannot be caged by the host.

Strategies for the creation of new molecular caging complexes depend on the application. For example, if a given host is considered for molecular shape sorting, one has to screen potential guests. In a dual manner, if a particular drug is considered for targeted delivery by a nanoscale carrier, one has to screen potential host molecules. Unfortunately, current experimental challenges hamper such high-throughput screening efforts and, in turn, make general synthetic approaches very time- and resource-demanding.[5] This issue clearly highlights the need for computational methods for the high-throughput screening of host and/or guest candidates prior to experimental validation.

In previous work, we proposed a computationally-efficient algorithm to predict if a given pair of molecules are likely to form a caging complex.[6] This algorithm takes two static molecular geometries of arbitrary shape as input; in other words, each molecule is represented by a three-dimensional union of balls of given radii, according to the classical hard-sphere model. Then, as our algorithm is based on a mathematically provable and conservative verification of the caging property, it predicts that a given host-guest pair forms a caging complex only when appropriate theoretical guarantees are met. Note that our caging verification algorithm was initially developed in the field of robotics (for applications to manipulation and path planning) where related concepts of caging were studied.

In this article, we present MoleQCage, a high-throughput screening tool for molecular caging prediction, based on our robotics-inspired caging verification algorithm.[6] MoleQCage takes as input a set of candidate host molecules and a set of candidate guest molecules. Then, for each pair of host-guest molecules, the underlying verification algorithm determines whether they are likely to form a caging complex, based solely on their geometries. For that, our algorithm considers uncertainties in the definition of molecular geometries and assesses the robustness of each caging prediction. Eventually, for each pair of candidate host-guest molecules, MoleQCage provides as output a prediction on whether it forms a caging complex (+) or not (-).

Caging Verification Algorithm

Given fixed conformations of a host and guest molecules, our algorithm uses an efficient representation of the (six-dimensional) configuration space of the guest to approximate its free space, i.e. the space in which the guest can move within the constraints imposed by the host.[6] A configuration of the guest molecule refers to its position and orientation in three-dimensional space. If the free space of the guest contains a bounded connected component (i.e. a finite-sized subspace in which every pair of configurations can be connected by a collision-free path), then we can prove that the guest is caged by the host.

In our method, molecular geometries are defined as unions of balls with atomic van der Waals radii, and uncertainties in these geometries are accounted for by varying the balls radii.[6] This is done by modifying all radii using a given Δr value, which is a parameter accessible to MoleQCage users. Varying these radii allows assessing the robustness of a caging prediction for a given host-guest pair by applying our caging verification algorithm to slightly different molecular geometries. Indeed, in cases where a host-guest pair might be predicted to form a caging complex based on given molecular geometries, small changes in these geometries might lead to a different prediction. In such cases, either because the guest can now escape or cannot fit the host cavity any more, we say that this host-guest pair forms a "weak" caging complex. In other cases, if the guest is consistently predicted to be caged by the host, we say that the host-guest pair forms a "strong" caging complex; if the guest is consistently predicted to not be caged by the host, we say that the host-guest pair do "not" form a caging complex. Therefore, for each evaluated pair of host-guest molecules, after applying the caging verification algorithm with radii perturbations based on a given Δr value (typically ± 0.3 Å) in MoleQCage, we can determine whether this pair of molecules (i) does "not" form a caging complex, (ii) forms a "weak" caging complex, or (iii) forms a "strong" caging complex.

Caging Prediction Use Cases

MoleQCage provides users with a flexible graphical user interface (GUI). To define molecular geometries, users can provide as input any file type containing atomic coordinates (such as mol2, pdb, or xyz). MoleQCage can then be applied to several caging prediction tasks.

Host-guest pairs screening with robustness assessment

MoleQCage can be used to screen a large number of guest molecules against a large number of host molecules. As the underlying caging verification is based on a geometric analysis, it is highly efficient, and therefore allows for such high-throughput screening. When users provide a set of candidate guests and a set of candidate hosts, MoleQCage runs our molecular caging verification algorithm, for all host-guest pairs of molecules, using multiple threads for computational efficiency. Based on the Δr value provided by users (or the default value), the robustness of these predictions can be assessed in MoleQCage, so that one can obtain a two-dimensional array with the values “weak”, “strong”, or “not”, for all host-guest pairs.

To illustrate this use case, we consider a set of four candidate guests (Figure 2), which are monohalobenzenes with relatively similar shapes and molecular volumes:[7] bromobenzene (BB), chlorobenzene (CB), fluorobenzene (FB), and iodobenzene (IB). Note that all crystal structures used in this work were obtained from the Cambridge Crystallographic Data Centre (CCDC) database. In addition, we consider a set of 38 candidate hosts (Figure 3), which are shape-persistent molecules with internal cavities. This list is a modified version of the CDB41 database,[8] from which duplicates have been removed and to which a few molecules have been added.[9,10]

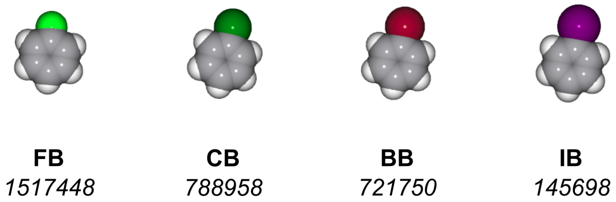

Figure 2: Monohalobenzenes evaluated as potential guests in the host-guest pair screening use case. Each column includes a molecular structure, an abbreviated name, and the identifier of the corresponding CCDC database entry. FB - fluorobenzene, CB - chlorobenzene, BB - bromobenzene, IB - iodobenzene.

In this scenario, using the default value for Δr (i.e., 0.3 Å), MoleQCage allowed us to efficiently screen all 152 host-guest pairs. Results show that 19 host-guest pairs are predicted to be strong caging complexes, 21 host-guest pairs are predicted to be weak caging complexes,

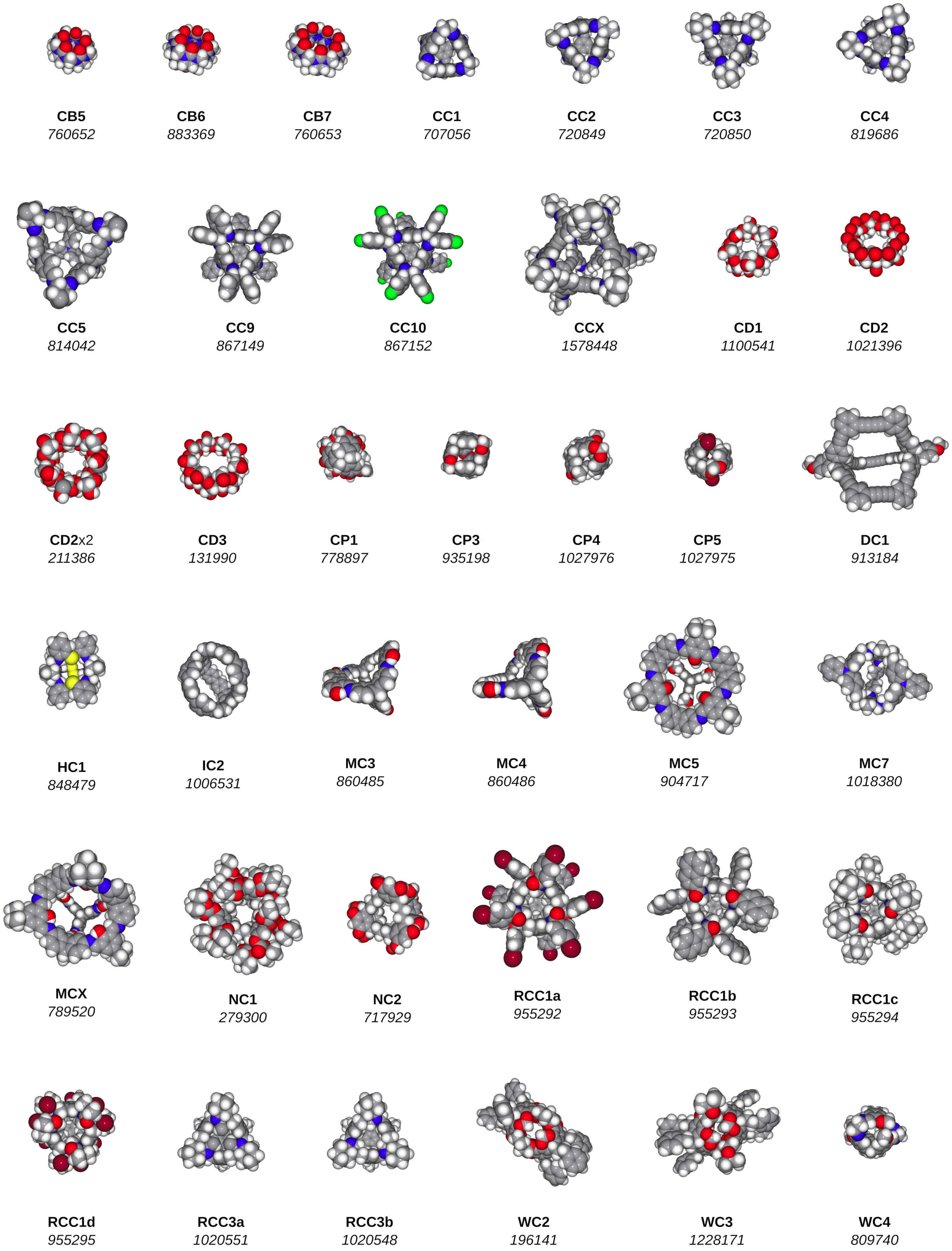

Figure 3: Set of 38 candidate hosts. For each candidate host, we provide a molecular structure, an abbreviated name, and the identifier of the corresponding CCDC database entry.

and 112 host-guest pairs are predicted not to form caging complexes (Table 1). Among the 19 strong caging complexes, 16 were formed by four hosts (CB6, RCC3b, WC2, and WC3) with the four guests. Three other hosts (CC3, NC1 and WC4) are predicted to form strong caging complexes, but only with FB (fluorobenzene, the smallest candidate guest), suggesting that their internal cavities are too small to fit larger candidate guests. Among all 38 candidate hosts, WC4 is the only one that is associated with the three possible outcomes, as it is predicted to form a strong caging complex with FB, to form a weak caging complex with CB and BB, and not to form a caging complex with IB. As a consequence, WC4 would be a good candidate for the separation of monohalobenzenes.

Table 1: Results produced by MoleQCage on the host-guest pair screening use case, which involved four candidate guests (Figure 2) listed in the columns, and 38 candidate hosts (Figure 3) listed in the rows. For each of the 152 host-guest pairs, the prediction is reported as: **strong** caging complex, **weak** caging complex, or **not** a caging complex.

host \ guest	FB	CB	BB	IB	host \ guest	FB	CB	BB	IB
CB5	not	not	not	not	DC1	not	not	not	not
CB6	strong	strong	strong	strong	HC1	weak	weak	not	not
CB7	not	not	not	not	IC2	not	not	not	not
CC1	not	not	not	not	MC3	not	not	not	not
CC2	not	not	not	not	MC4	not	not	not	not
CC3	strong	weak	weak	weak	MC5	not	not	not	not
CC4	weak	not	not	not	MC7	not	not	not	not
CC5	not	not	not	not	MCX	not	not	not	not
CC9	weak	weak	weak	weak	NC1	strong	weak	weak	weak
CC10	not	not	not	not	NC2	weak	weak	weak	weak
CCX	not	not	not	not	RCC1a	weak	not	not	not
CD1	not	not	not	not	RCC1b	weak	not	not	not
CD2	not	not	not	not	RCC1c	not	not	not	not
CD2x2	not	not	not	not	RCC1d	not	not	not	not
CD3	not	not	not	not	RCC3a	not	not	not	not
CP1	not	not	not	not	RCC3b	strong	strong	strong	strong
CP3	not	not	not	not	WC2	strong	strong	strong	strong
CP4	not	not	not	not	WC3	strong	strong	strong	strong
CP5	not	not	not	not	WC4	strong	weak	weak	not

Hosts or guests screening with implicit molecular flexibility

As our caging verification algorithm analyzes static conformations of molecules, it does not explicitly account for molecular flexibility. However, MoleQCage allows accounting for implicit molecular flexibility. For that, instead of providing a set of different hosts (or guests), users can provide a set of conformations for a single host (or guest). These conformations can be obtained from structural databases (such as the Protein Data Bank) or via molecular simulations, such as molecular dynamics (MD). Therefore, users can screen a set of guest candidates against multiple conformations of a given host molecule, or conversely screen a set of host candidates against multiple conformations of a given guest molecule.

To illustrate this use case, we consider a set of three candidate guests (Figure 4): mesitylene (Mes), m-xylene (mX), and 4-ethyltoluene (4ET). As host molecule, we consider CC3 (Figure 3) and use a conformational ensemble containing 515 conformations produced by an MD simulation reported in related work.[11]

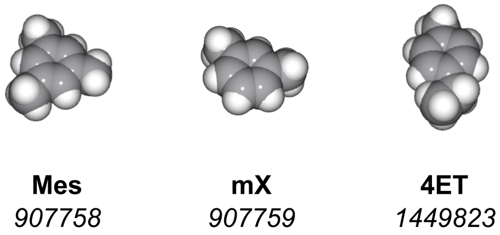

Figure 4: Molecules evaluated as potential guests in the guest screening use case involving host flexibility. Each column includes a molecular structure, an abbreviated name, and the identifier of the corresponding CCDC database entry. Mes - mesitylene, mX - m-xylene, 4ET - 4-ethyltoluene.

In this scenario, using a value of 0.1 Å for Δr, MoleQCage allowed us to efficiently screen all 1,545 host-guest pairs. Results show that CC3 forms a strong caging complex with Mes, and forms a weak caging complex with mX, but do not form a caging complex with 4ET (Figure 5). Indeed, 323 out of 515 CC3 conformations (63%) were predicted to form a strong caging complex with Mes; 408 out of 515 CC3 conformations (79%) were predicted to form a weak caging complex with mX; and 340 out of 515 CC3 conformations (66%) were predicted to not form a caging complex with 4ET. This is in agreement with experimental results reported for these host-guest candidates, which showed that 4ET could easily travel through CC3's windows, that mX could escape CC3's cavity but not as easily as 4ET, and that Mes was properly caged by CC3.[12]

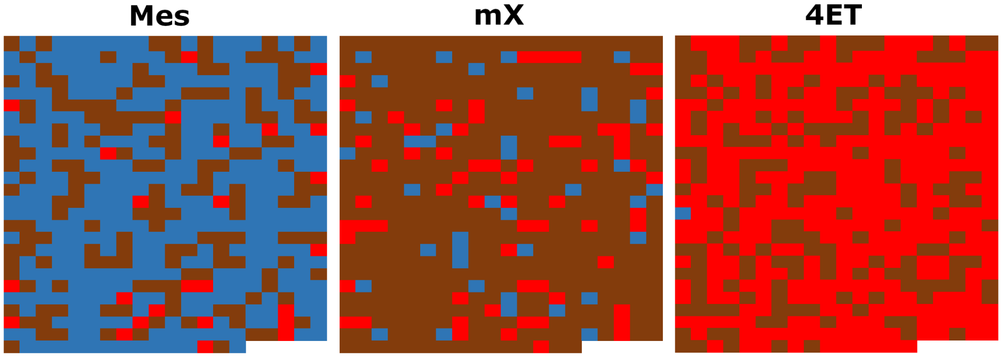

Figure 5: Results produced by MoleQCage on the guest screening with host flexibility use case. This scenario involved three candidate guests called Mes, mX, and 4ET (Figure 4) and 515 conformations of the candidate host CC3. For each host conformation, the prediction is reported as: a blue square for a strong caging complex, a brown square for a weak caging complex, or a red square if this is not a caging complex.

Caging prediction for a host-guest pair with implicit molecular flexibility

Users can restrict their analysis to a single host-guest pair and implicitly consider the flexibility of both molecules by providing a set of host conformations and a set of guest conformations. As in other use cases, MoleQCage will allow users to produce a two-dimensional array containing the values "weak", "strong", or "not", for all conformation pairs. Since the caging verification algorithm is computationally efficient, it is totally realistic to consider screening a large number of host/guest conformations, and therefore obtain a caging prediction almost as accurate as if molecular flexibility was explicitly modeled.

To illustrate this use case, we evaluate 4ET (Figure 4), using four manually-generated conformations, against CC3, using the 515 MD conformations mentioned in the previous section. In this scenario, using a value of 0.1 Å for Δr, MoleQCage allowed us to efficiently screen all 2,060 host-guest pairs. Results are very similar to what was obtained without considering the flexibility of 4ET because this small molecule has only one rotatable bond (Figure 6). Therefore, the prediction is still that 4ET and CC3 do not form a caging complex.

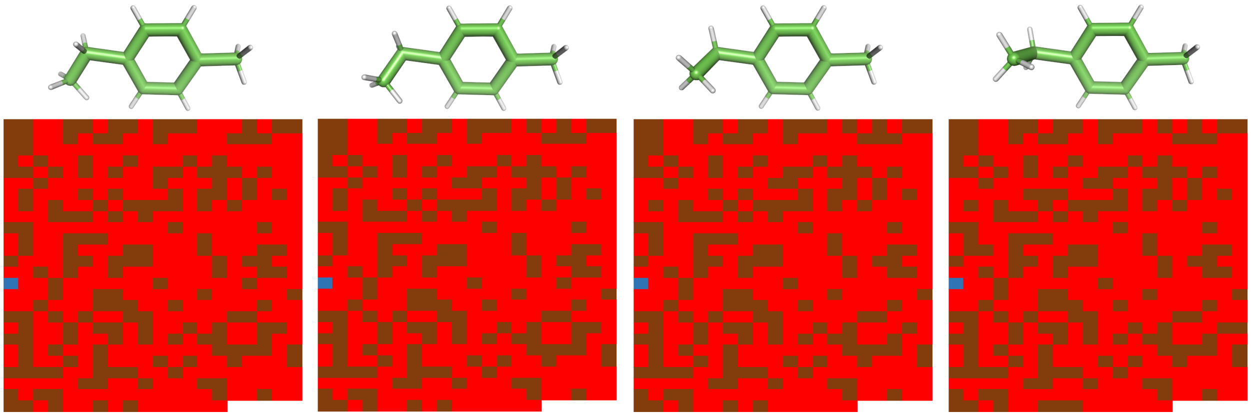

Figure 6: Results produced by MoleQCage on the caging prediction with molecular flexibility use case. This scenario involved four conformations of the candidate guest 4ET and 515 conformations of the candidate host CC3. For each host conformation, the prediction is reported as: a blue square for a strong caging complex, a brown square for a weak caging complex, or a red square if this is not a caging complex.

Data and Software Availability

MoleQCage is distributed as Linux-based software with a graphical user interface. It is available online free of charge at `https://hub.docker.com/r/dantrigne/moleqcage` in the form of a Docker container. Documentation and examples are available as Supporting Information and online at `https://hub.docker.com/r/dantrigne/moleqcage`. Data used to test and validate MoleQCage are provided as Supporting Information.

Supporting Information Available

The following files are available free of charge at the ACS website:

- supplement.pdf: instructions on how to install and use MoleQCage;

- molecules.zip: mol2 files of molecules used to test and validate MoleQCage;
- grids.zip: rotation grids used to decompose the space of all possible rotations.

Acknowledgement

This research was supported by the Knut and Alice Wallenberg Foundation [A.K., A.V., D.K.]; the Swedish Research Council [A.K., A.V., D.K.]; the Rice University Funds [L.K.]; and the Medical Research Council (MRC) [MC_UU_00009/2 to D.D.]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

(1) Ahmad, N.; Younus, H. A.; Chughtai, A. H.; Verpoort, F. Metal-organic molecular cages: Applications of biochemical implications. *Chem. Soc. Rev.* **2015**, *44*, 9–25.

(2) Bhaskar, S.; Lim, S. Engineering protein nanocages as carriers for biomedical applications. *NPG Asia Mater.* **2017**, *9*, e371.

(3) Sigl, C.; Willner, E. M.; Engelen, W.; Kretzmann, J. A.; Sachenbacher, K.; Liedl, A.; Kolbe, F.; Wilsch, F.; Aghvami, S. A.; Protzer, U.; Hagan, M. F.; Fraden, S.; Dietz, H. Programmable icosahedral shell system for virus trapping. *Nat. Mater.* **2021**, *20*, 1281–1289.

(4) Mastalerz, M. Porous shape-persistent organic cage compounds of different size, geometry, and function. *Acc. Chem. Res.* **2018**, *51*, 2411–2422.

(5) Greenaway, R. L.; Santolini, V.; Bennison, M. J.; Alston, B. M.; Pugh, C. J.; Little, M. A.; Miklitz, M.; Eden-Rump, E. G. B.; Clowes, R.; Shakil, A.; Cuthbertson, H. J.; Armstrong, H.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. I. High-throughput

discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.* **2018**, *9*, 2849.

(6) Kravchenko, O.; Varava, A.; Pokorny, F. T.; Devaurs, D.; Kavraki, L. E.; Kragic, D. A robotics-inspired screening algorithm for molecular caging prediction. *J. Chem. Inf. Model.* **2020**, *60*, 1302–1316.

(7) Atwood, J., Ed. *Encyclopedia of Supramolecular Chemistry*; Marcel Dekker, 2004.

(8) Miklitz, M.; Jiang, S.; Clowes, R.; Briggs, M. E.; Cooper, A. I.; Jelfs, K. E. Computational screening of porous organic molecules for xenon/krypton separation. *J. Phys. Chem. C* **2017**, *121*, 15211–15222.

(9) Mastalerz, M.; Schneider, M. W.; Oppel, I. M.; Presly, O. A salicylbisimine cage compound with high surface area and selective CO_2/CH_4 adsorption. *Angew. Chem. Int. Ed. Engl.* **2011**, *50*, 1046–1051.

(10) Pugh, C. J.; Santolini, V.; Greenaway, R. L.; Little, M. A.; Briggs, M. E.; Jelfs, K. E.; Cooper, A. I. Cage doubling: Solvent-mediated re-equilibration of a [3 + 6] prismatic organic cage to a Large [6 + 12] truncated tetrahedron. *Cryst. Growth Des.* **2018**, *18*, 2759–2764.

(11) Miklitz, M.; Jelfs, K. E. Pywindow: Automated structural analysis of molecular pore. *J. Chem. Inf. Model.* **2018**, *58*, 2387–2391.

(12) Mitra, T.; Jelfs, K. E.; Schmidtmann, M.; Ahmed, A.; Chong, S. Y.; Adams, D. J.; Cooper, A. I. Molecular shape sorting using molecular organic cages. *Nat. Chem.* **2013**, *5*, 276–281.