Learning Feasibility, Shape and Cavity of Metal-Organic Cages from their Building Blocks

Simone Zucchelli [1], **Fernanda Duarte** [2], **Pascal Friederich** [3], **Paola Posocco** [1, †]

[1] Department of Engineering and Architecture, University of Trieste, Via Alfonso Valerio 6/1, 34127 Trieste, Italy
[2] Chemistry Research Laboratory, Department of Chemistry, University of Oxford, 12 Mansfield Road, OX1 3TA Oxford, United Kingdom
[3] Institute of Nanotechnology (INT), Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany
corresponding author's email: *paola.posocco@dia.units.it*

Purpose of the work. Metal-organic cages (MOCs) are discrete, hollow self-assemblies of metal ions and organic ligands (building blocks), with applications in molecular recognition, catalysis and separation.[1] Predicting whether a set of designed building blocks will form a cage, and determining its resulting shape and cavity, typically requires synthetic efforts and costly quantum-chemical calculations. Here we overcome these limitations by predicting MOC structure and features directly from their simple building blocks.

Methods. Using the CSD Python API, ASE and NetworkX, we mined the Cambridge Crystallographic Data Centre (CCDC) to curate an open dataset of experimentally known MOCs.[2] Each cage was encoded as a graph of metals and ligands, using PyTorch and PyTorch Geometric, with cavity properties calculated via CageCavityCalc.[3] Structural penalty, i.e. the energetic cost of distorting ligands upon assembly, was estimated on geometries optimised at GFN2-xTB level.[4] A first model learns individual ligand penalty within the MOC and transfers it to two models requiring only a single ligand geometry: the first uses the relaxed ligand to screen building-block libraries before cage assembly, and the second uses cage-bound ligand geometry to characterize a specific system. Then, a graph neural network (GNN) combines these ligand features with metal–ligand connectivity to predict the whole assembly properties.

Results. Trained in a multi-task scheme, our top three models predict MOC structural feasibility outperforming descriptor-based and graph-only models on regression tasks (structural feasibility and cavity properties). On polyhedral shape classification, they reach an accuracy of 0.9, matching the random forest baseline performance. More importantly, differently from the baseline, they achieve that using only the building blocks rather than the full assembled cage.

New or breakthrough work to be presented. (a) A curated, machine learning-ready open dataset of experimentally known MOCs; (b) three GNN models for discovery and characterization of novel MOCs; (c) their combination into an extensible, and reproducible pipeline.

Conclusions. Representing a cage as a graph of metals and ligands captures it better than aggregated descriptors. Within the framework developed, structural feasibility, polyhedral shape and cavity properties can be predicted without quantum-chemical calculations on the cage itself, supporting both the discovery of promising building-block combinations for synthesis as well as the rationalisation of existing cages.

References

[1] a) C. García-Simón, M. Costas and X. Ribas, Chem. Soc. Rev. 45, 40 (2016). b) Y. Fang, J. A. Powell, E. Li, Q. Wang, Z. Perry, A. Kirchon, X. Yang, Z. Xiao, C. Zhu, L. Zhang, F. Huang and H.-C. Zhou, Chem. Soc. Rev. 48, 4707 (2019).
[2] C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, Acta Crystallogr. B 72, 171 (2016).
[3] V. Martí-Centelles, T. K. Piskorz and F. Duarte, J. Chem. Inf. Model. 64, 5604 (2024).
[4] C. Bannwarth, S. Ehlert and S. Grimme, J. Chem. Theory Comput. 15, 1652 (2019).