Understanding MOF nucleation from solution with Evolving Graphs

Loukas Kollias[1], Roger Rousseau[1], Vassiliki-Alexandra Glezakou[1*], Matteo Salvalaglio[2*]

[1]Basic and Applied Molecular Foundations, Pacific Northwest National Laboratory, Richland, WA, 99352 United States; [2]Thomas Young Centre and Department of Chemical Engineering, University College London, London WC1E 7JE, United Kingdom

KEYWORDS: Molecular Dynamics, Metal – Organic Frameworks, Graph Theory

ABSTRACT: Molecular modeling is ordinarily employed to understand the synthesis of complex materials. In this work, we investigate the collective assembly of building units that have been experimentally observed to initiate Metal-Organic Framework (MOF) nucleation. MOFs exhibit attractive characteristics such as remarkable surface area and diverse porosities, however, a mechanistic understanding of their synthesis and scale-up remains underexplored due to the complicated nature of the building block interactions. Here, we tackle this problem with large-scale molecular dynamics simulations under a variety of synthesis conditions and mixture compositions. We observe that the connectivity of building units, as well as their level of crystalline order and fractal dimension, largely vary depending on the synthesis conditions. However, these properties naturally emerge when interpreting the self-assembly process of MOF nuclei as the time-evolution of an undirected graph. The results show that solution-induced conformational complexity and ionic concentration have a dramatic effect on the morphology of clusters emerging during assembly, such diversity is captured by key features of the graph representation. Principal Component Analysis (PCA) on graph properties successfully deconvolutes MOF self-assembly to be characterized by a small number of molecular descriptors, such as average coordination number between half-secondary building units (half-SBUs) and fractal dimension, which can be followed by time-resolved spectroscopy. We conclude that graph theory can be used to understand complex processes such as MOF nucleation by providing molecular descriptors accessible by both simulation and experiment.

INTRODUCTION

Metal-Organic Frameworks (MOFs) are a class of porous materials that thanks to their high porosity and surface area, have ignited interest in a plethora of applications including carbon capture and storage[1, 2], separations[3], extraction of water from air[4], electrodes[5] and drug delivery[6]. Nevertheless, the stability of MOFs is lower than other porous materials and their scaling-up is still problematic, thus reducing their applicability[7, 8]. The presence of defects in MOFs is known to affect their thermomechanical properties, stability, synthesis costs, and overall suitability for industrial applications[9, 10].

This has lead to recent efforts to understand the detailed mechanisms associated with MOF synthesis in order to regulate the extent of defects[11, 12]. The formation of secondary building units (SBUs) during the early stages of synthesis is crucial in determining the final properties of a MOF. Ferey et al.[13] suggested a synthesis mechanism involving the formation of pre-nucleation building units (PNBUs) and their subsequent nucleation. These are soluble zero-charged species such as the half-SBUs mentioned in this work. More experimental works have identified PNBUs and evaluated their role in the final MOF structure following the approach of synthesis through SBU formation[14-18]. Recently, Liu et al.[19] suggested a three-step MOF nucleation mechanism through a mixed experimental and computational work. They identified metastable structures that recrystallize into the MOF, hence acting as precursors to the nucleation of crystalline MOFs, but could not elucidate the molecular mechanisms governing the process.

Computational studies on the early stages MOF self-assembly are rather limited[20]. Yoneya et al.[21] studied MOF self-assembly with a focus on optimizing simulation parameters using dummy atoms in implicit solvent. Biswal and Kusalik[22] have studied MOF self-assembly using Langevin dynamics, and their results imply the existence of several local energy minima associated with the process. Wells et al.[23] investigated the early stages of MOF synthesis using Monte Carlo methods and developed an algorithm able to distinguish between different phases with respect to composition. Colon et al.[24] focused on the self-assembly of MOF-5 using enhanced sampling methods, but they did not consider all relevant metastable states as they restrained the endpoints of the biased simulation. Cantu et al.[25] investigated the assembly of MIL-101(Cr) building blocks at the density functional theory level, and identified possible SBU isomers through modelling kinetics of formation. Finally, the formation of "metal-free" covalent organic frameworks has been studied using coarse-grained models for building blocks[26, 27]. The shape and structure of clusters can be exam-

ined as a means to characterize the growth of complex materials. In this respect, the fractal dimension is a well-known descriptor of compactness used to describe the self-assembly of metal-ligand systems[28]. Recently, new synthetic strategies to form self-similar MOFs that exhibit fractal geometry were suggested[29, 30]. Maurer et al.[31] observed that fractal dimension provides useful insight into the spatial arrangement of structures which have similar radii of gyration. Tsao et al.[32] used the fractal dimension to explore the potential of hydrogen storage inside MOF pores. Goesten et al.[33] investigated MOF growth through characterizing the fractal nature of precursors. At last, Liu et al.[19] linked the shape of clusters with crystallinity during MOF nucleation.

Graph theory has been used to model nucleation of metal-oxide compounds[34], or construction of molecular polyhedra[35]. Also, it has been successfully applied to study the properties of metal-ligand systems[36] and MOFs[37-39], and characterize the dynamic behavior of functional materials[40]. In the case of MOF self-assembly, a dimensionality reduction based on graphs is a computationally efficient way to analyze simulation results, as a MOF can be naturally thought of as a network of nodes, i.e. the metal centres, connected by edges, i.e. the ligands.

In a previous work, we have demonstrated how the initial population of isomer half-SBUs, the choice of solvent and the concentration of ionic species can affect the formation of defects based on interactions between couples of half-SBUs[41]. In this work we use large scale atomistic simulations of the self-assembly of MOF precursors from preformed SBUs in order to systematically understand how the composition of the environment and the relative abundance of different half-SBUs isomers affects the collective assembly behavior that leads to nucleation. We perform an unsupervised clustering analysis of half-SBUs based on a graph-based model[42] in order to identify MOF precursors emerging from solution, to calculate the properties of their interconnected structures.

Finally, we carry out a principal component analysis to identify the properties which largely determine how self-assembly proceeds under various conditions. This way we deconvolute the characterisation of the complex MOF self-assembly process by projecting various properties on the low-dimensional space of principal components. This allows us to evaluate various solution compositions and group them based on the similarity of the resulting assembly mechanism; thus, offering a mechanistic understanding of the early stages of MOF self-assembly.

METHODS

Simulation setup. Molecular Dynamics (MD) simulations were performed in explicit solvent. Water has been represented with the TIP3P model[43] and ions with the OPLS-AA force field[44]. Our analysis is carried out, apart from water, in N,N-dimethylformamide (DMF) using force field parameters compatible with OPLS-AA[44] obtained from the virtual-chemistry.org database[45, 46]. The MOF half-SBUs were modelled using force field parameters obtained from a previous work[41]. The leapfrog integrator was used to propagate dynamics of the system with a time step of 2 fs. The LINCS[47] algorithm preserved the distances of bonds involving hydrogen atoms. The cut-off for non-bonded interactions is 10 Å. Long range electrostatics were treated using the Particle-Mesh Ewald (PME)[48] scheme. The Bussi-Donadio-Parrinello thermostat[49] and the Berendsen barostat[50] preserved the temperature and pressure at 493 K and 3,500 bar respectively for an equilibration period of 5 ns. Production molecular dynamics simulations followed using the Parrinello-Rahman Barostat[51] with a relaxation time of 2 ps in water and 4 ps in DMF. Production simulations were carried out for 100 ns. The system consists of 132 half-SBUs in explicit solvent. The SBU are introduced indifferent proportions of three different Metal-Ligand (ML) isomers each with different orientations of the ligands about the Cr_3O_3 core denoted as, MLA, MLB, and MLC[25, 41]. This results in a total number of approximately 800,000 atoms in water and 500,000 atoms in DMF initially dispersed in a cubic box with an edge of 200 Å. Two initial compositions of half-SBUs are considered. In the first one (purely MLA), there are 132 MLA half-SBUs in solution. The second one (equiprobable MLA, MLB, and MLC), consists of 44 MLA, 44 MLB and 44 MLC in solution. Simulations were performed with periodic boundary conditions in three dimensions, using GROMACS 2018[52]. Chemical structures were visualized with VMD[53].

Distance and adjacency matrices. At first, a (n x n) distance matrix is constructed, where n is the number of the half-SBUs. The generic ij element of the matrix corresponds to the Euclidean distance between the central oxygen atoms of the i^{th} and j^{th} half-SBUs units was used as the argument of a step function to define the adjacency of two half-SBUs . A cutoff distance of 15 Å is chosen as this value lies between the first and second coordination shells emerging from the calculation of the pair radial distribution function between central oxygen atoms of half-SBUs. In the adjacency matrix, the element ij is equal to the unity if the distance between i and j is below the cutoff otherwise it is set to zero. Examples of the distance and adjacency matrices are discussed in the Supporting Information (SI), section I.

Graph-based clustering. The clustering analysis of half-SBUs is performed by converting molecular structures into lower dimension graph representations. We consider the central oxygen atoms of the building units (nodes) and connect these nodes advised by the dissimilarity matrix formed through a distance criterion. Then, we identify clusters as connected components in the graph, where the smallest cluster consists of two half-SBUs (dimer).[54] Departing form depth first search (DFS)[54] that can be used to identify clusters, we analyze the properties of clusters as components of a graph; hence enrich the information we have for each cluster. Also, we calculate the mass of each cluster as the sum of the masses of its constituent particles. Furthermore, the local environment of each half-SBU is characterized by enumerating the neighbors of each node. This is a measure of the coordination of half-SBUs in the cluster. Also, we calculate the number of "free" half-SBUs as isolated nodes, without any edges attached. This allows us to further calculate certain properties of the graph, such as the number of connections (average neighborhood degree of each node,

named graph degree for simplicity), the extent of interconnected triplets (transitivity) and the number of nodes which connect with similar ones based on their degree (assortativity coefficient). The graph representation was constructed using the NetworkX Python library[42].

Spherical radius and radius of gyration. The spherical radius and the radius of gyration are calculated for each cluster. The first follows from the ideal process of including each cluster in a sphere. The radius of that sphere would then be half the maximum distance between any two metal centers in this cluster. The radius of gyration[55] provides insight into the distribution of mass in complex structures and is used to calculate the fractal dimension of each cluster. Periodic boundary conditions in three dimensions are appropriately considered in all these calculations.

Fractal Dimension. The fractal dimension D_f was computed using a power law approach[56, 57] in the form $M = R_{gyr}{}^{D_f}$.

Where M is the cluster mass, and R_{gyr}its mass-weighted radius of gyration. In single structures we consider all atoms within a spherical volume extending from the center of mass of a cluster. For all atoms included into spherical volumes of increasing radii, we calculate the total mass and the mass-weighted radius of gyration. The fractal dimension is then calculated as the angular coefficient of the linear function obtained by fitting $\log(R_{gyr})$ against log(M).

This is repeated for all clusters consisting of at least three half-SBUs. We consider spherical shells ranging from the minimum distance between centers of half-SBUs in the cluster increased by 1 Å, to the maximum distance increased by 5 Å. This allows to have enough atoms to calculate a meaningful radius of gyration for all volumes, with the largest radius containing the entire cluster.

A similar analysis is applied to clusters that spontaneously emerge during simulation. We observe the growth of a fractal pattern by plotting the size of clusters against the corresponding radii of gyration. Then, we use the power-law relationship to calculate the fractal dimension of clusters formed along the simulation trajectory as in using the power law approach. In this case, however, we substitute the total mass M with the number of half-SBUs units in the cluster. For non-mass-weighed R_{gyr} of the cluster is computed from the positions of the centers of mass of the units.

Diffraction patterns. We calculate the X-Ray powder diffraction (XRD) patterns of clusters emerging throughout the simulation using PyMatGen[58, 59]. The wavelength of the Z-Ray source used is equal to Cuk-α radiation λ=1.54184 Å. We consider angles that differ less than 10^{-5} radians to have the same intensity. We scale intensities so that the unity is the maximum value and scaled intensities less than 10^{-5} were considered to be negligible. Hydrogen atoms are not present in the calculation to be consistent with the pattern of the experimental crystal structure. This analysis is performed for the five largest clusters at every ns of the simulation. Consequently, distance metrics can be used to evaluate the similarity of each structure with the reference[60, 61]. In this effort, we used the various distance metrics such as the Euclidean, Hellinger, cosine, Manhattan, χ^2 and Canberra distances[62]. Similarity is calculated as the difference of the distance from the unity[62]. It should be noted that distance is normalized by its maximum value in all simulations to calculate similarity. Zero distance means highly similar, while distance equal to the unity means patterns are not similar to MIL-101(Cr)[63].

Effect of ions and solvent. Ions (Na^+, F^-) are added at concentrations of 0.035M in water and 0.075M in DMF. We chose small concentrations as it was suggested that this is the optimal balance between crystallinity and salt precipitation[41]. Furthermore, a different solvent (DMF) can significantly affect the energetics of conformational transformations of the solute[41, 64]. In this context, we can assess the effect of guest molecules on the collective assembly of half-SBUs. Finaly, there is experimental evidence that studying the solution in the early stages of assembly can significantly improve our understanding of the mechanism of MOF synthesis[65].

Calculation of rates. We calculate the rate of cluster formation during self-assembly following the approach discussed by Yuhara et al.[66]. In more detail, we calculate the rate of cluster formation per unit volume, J. This is computed directly from unbiased simulations by estimating the partial derivative of the number of clusters formed, N(t), with respect to simulation time, t, and then normalizing it by the volume of the simulation box V:

$$J = \frac{1}{V} \cdot \frac{\partial N(t)}{\partial t} \qquad (2)$$

In the expression above, the volume of the system shows negligible fluctuations during the simulation, hence it is considered a constant. A variety of methods is employed to calculate the partial derivative. At first, we use a linear fit of the transient period of the simulation and then we calculate J from the slope. Also, we approximate J as the ratio of the differences in N(t), t between consecutive time points, namely $\Delta N(t), \Delta t$ as implemented in the NumPy software package[67]. At last, we use the more accurate finite differences method discussed by Fornberg[68], also implemented in NumPy[67]. Ultimately, the rate is calculated using the average value of the derivative in the transient period. Differences arising due to the use of various methods to calculate rates are accounted for as the standard deviation of the estimate of the assembly rate in each simulation. An example calculation can be found in the SI, section VI.

Principal Component Analysis. We calculate different sets of data which characterize our analysis of the early stages of MOF self-assembly. A dimensionality reduction can be performed while retaining meaningful information by projecting data on principal components that possess most of the variation of the dataset[69]. In this effort, we consider the average values of quantities corresponding to the largest cluster along the trajectory of each simulation. We normalize data to have a mean of zero and standard deviation equal to unity. This is done to combine data that have different units and magnitude[70]. Then, we identify eigenvectors and eigenvalues of the covariance matrix[70]. At last, we project data on the eigenvectors with the highest eigenvalues (principal components).

RESULTS

Simulations in pure water. We start our analysis with the system of MLA half-SBUs in pure water. Half-SBUs gradually attach to one large cluster leaving no "free" half-SBUs after approximately 70 ns. This cluster is highly ordered and forms pores. The Q interaction between half-SBUs[64], is prevalent in this case. This interaction type is which is entropically favoured in water and features π-π stacking interactions.

We observe a rapid decrease in the number of clusters in the first 10 ns as almost half-SBUs are already connected with other two neighbors at this time. Also, we monitor the size of the five largest clusters. After approximately 70 ns the clusters start sintering, incorporating into the large cluster during this time. On average, each half-SBU that belongs to this cluster has three other neighboring half-SBUs. We believe that this high degree of interconnectivity is what holds this cluster intact for the rest of the simulation. The five largest clusters formed of MLA half-SBUs in pure water at the very early stages of the simulations are reported in **Figure 1** along with their corresponding graph representations.

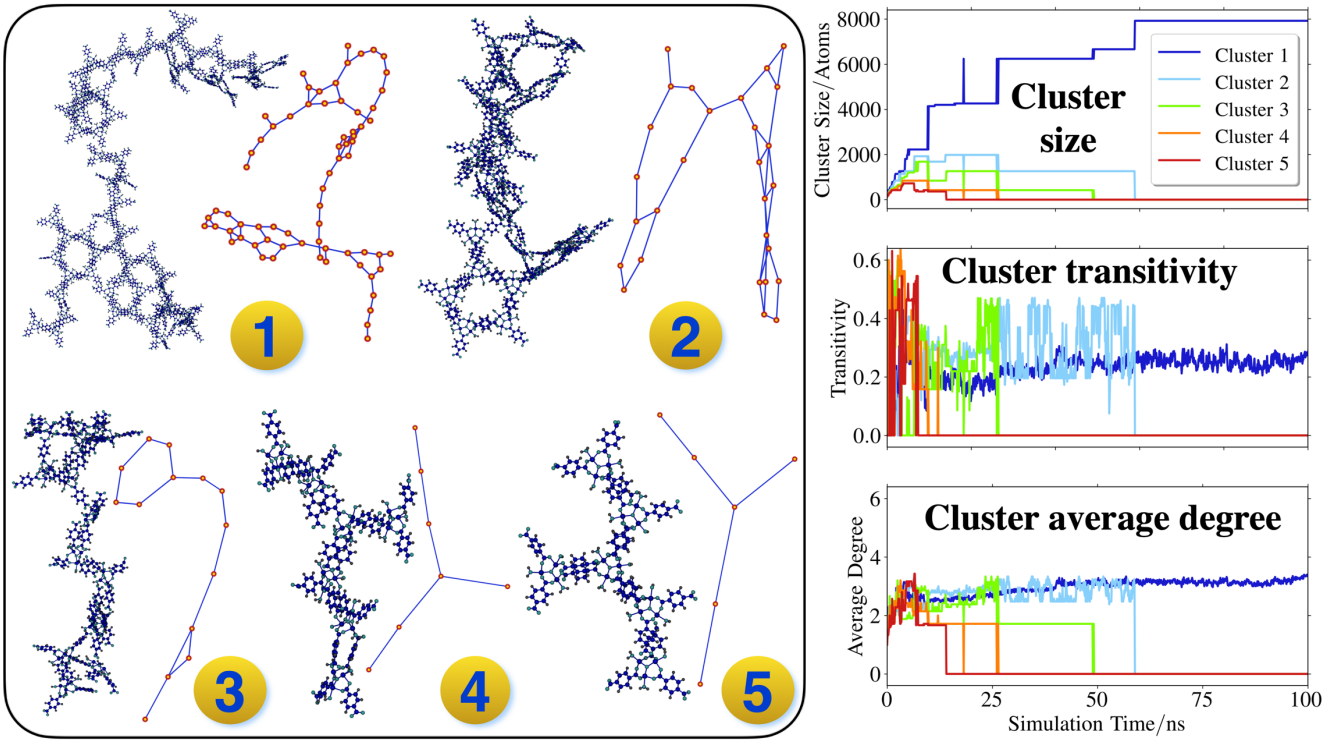


Figure 1. Purely MLA half-SBUs in water. Left panel: molecular (left) and graph (right) representations the five largest clusters (ranked by size) after 5 ns of the production simulation. Atom color code: Cr – lime, O – cyan, H – gray, C – blue. Right panel: Trajectory of the size, transitivity and average degree of the five largest clusters throughout the simulation. A larger representation of molecular structures is available in the SI, section I. Units are separated from the axis title by a forward slash.

An initially equiprobable distribution of half-SBUs in water results in the formation of two relatively smaller clusters. Also, these are less ordered than the previous case where MLB and MLC were absent. Nevertheless, the clusters present higher dimensionality than the cluster emerging from the purely MLA system.

The number of clusters is gradually decreasing to 2, reaching a plateau after 50 ns. One of the clusters is almost 30% larger than the other (4,380 and 3,420 atoms respectively). The average number of neighbors each half-SBU has is the largest of all cases as it plateaus to a value greater than 5 after 50 ns. At this point it should be noted that the figures resulting from the analysis of clusters for all mixture compositions studied in this work are available in the SI, section II. Additionally, the probability density of different cluster sizes is discussed in the SI, section III. Structures emerging from the simulation of assembly in pure water along with the corresponding graph representations after 20 and 100 ns of the production simulations for both "purely MLA" and equiprobable half-SBU distributions are shown in **Figure 2**. The structures and graph representations for the rest of the simulations are available in the SI, section II.

Furthermore, we used two clusters formed by MLA isomers in pure water after 100 ns and ran a relatively short MD simulation in vacuo. The reason is to evaluate whether clusters of higher dimension can be formed from 2D sheets, as the MIL-101 crystal is a three-dimensional network[71]. A cluster of increased fractal dimension emerged; hence such clusters could possibly develop after longer times of self-assembly. This is further discussed in the SI, section IV.

Simulations in water with ions. The introduction of ions (Na^+, F^-) considerably affects the dynamics of assembly. Ions promote numerous small clusters in contrast with pure water. Crystal-like SBUs are formed in the presence of

ions;[41] hence ions can help in healing defects during assembly. Also, small clusters of higher dimensions forming in this case as it is desirable[71]. At last, we observe small clusters forming during assembly under both half-SBU distributions; hence solute composition does not significantly affect the early stages of self-assembly in water with ions.

In further detail, the number of clusters gradually reaches a plateau at 25 ns for MLA half-SBUs, while this happens much faster in the equiprobable half-SBU distribution. There are approximately 13 and 25 clusters in the "purely MLA" and the "equiprobable half-SBUs" cases respectively. The two largest clusters consist of 1,320 and 1,140 atoms when there are purely MLA half-SBUs and 780 and 480 atoms when there is an equiprobable distribution of half-SBUs.

The average number of half-SBU neighbors is close to 3.1 on average in the "purely MLA" case, and only slightly lower (close to 2.4 on average) in the equiprobable half-SBU distribution. We note that the number of neighbors in the "purely MLA" scenario is similar to the one in pure water; hence ions do not alter the degree of connectivity in this case. A different behavior is observed in the equiprobable distribution of half-SBUs as ions tend to considerably decrease the number of half-SBU neighbors. Therefore, the coordination within the clusters is highly affected by the presence of ions in water when only MLB and MLC are considered. The choice of a small concentration of ions is further validated by results obtained from a larger concentration as discussed in the SI, section V.

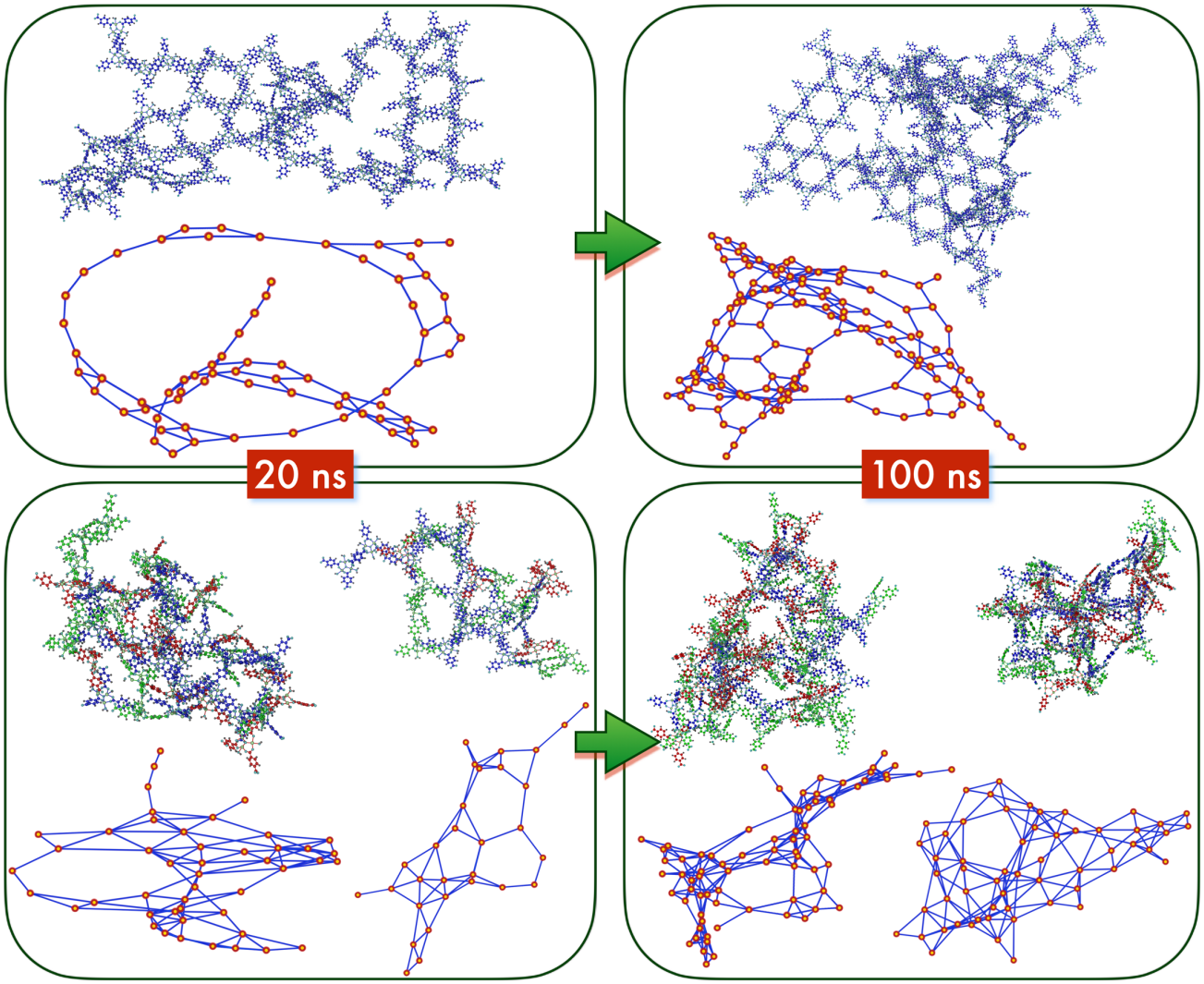


Figure 2. Clusters emerging from self-assembly in pure water. Molecular structures formed of the "purely MLA" (top) and "Equiprobable" (bottom) systems after 20 ns (left) and 100 ns (right) production simulations are shown along with the respective graph representations. Atom color code: Cr – lime, O – cyan, H – gray, C – blue (MLA), red (MLB), green (MLC). Nodes in the graph represent the central oxygen atom of each monomer and an edge is drawn between connected nodes. Graph color code: nodes – gold, edges – blue.

Simulations in DMF. Solvent effects are investigated through simulation of the early stages of MOF self-assembly in DMF. The system, consisting of purely MLA half-SBUs, rapidly forms a large cluster where the Q configuration prevails, and linear chains are also formed. Pores form in a similar fashion as in water. In the equiprobable

half-SBUs case, we see one long cluster forming in contrast with corresponding simulations in water. Consequently, DMF leads to a decrease in the dimensionality of the resulting structure in the equiprobable half-SBUs scenario.

The number of clusters is rapidly decreased when purely MLA half-SBUs are present in DMF. Equiprobable half-SBUs lead to the formation of two large clusters (4,080 and 3,420 atoms respectively) between 30 and 50 ns. The average number of neighbors plateaus at around 3 when purely MLA half-SBUs are present, while it is slightly higher when MLB and MLC are introduced.

Simulations in DMF with ions. Ionic species show a completely different behavior in DMF when compared with water. The "purely MLA" system forms a large, ordered, and high-dimensional cluster compared with the one formed in absence of ions. A similar behavior is observed when ions are added in the "equiprobable half-SBUs" scenario.

The time evolution of the number of clusters shows a similar trend in both half-SBU distributions resulting in 3-4 clusters after 100 ns. In "purely MLA", we observe a slightly higher degree of interconnectivity as the number of neighbors is close to 3.3 against 3 in the "equiprobable half-SBUs" scenario.

Graph analysis. A graph representation of the system allows us to calculate of certain properties of the molecular network in every simulation. The equiprobable distribution of species results in higher numbers of connections per half-SBU, hence half-SBUs interact more than when MLA is the only species present. Also, the equiprobable distribution of species and the presence of ions leads to higher transitivity values; hence networks are more tightly connected as more triplets of interconnected members exist. In contrast, purely MLA systems have higher assortativity coefficients, except in the presence of ions in DMF. Consequently, MLA leads to assortative mixing where similar molecules (which have the same number of connections on average) are interconnected more.

Overall, larger networks, formed when purely MLA isomers are present, feature members that connect to others with whom they share similarities (e. g. of the same number of connections), but they are not as strongly connected as smaller clusters emerging in the equiprobable distribution of species. The latter "communicate" more with others that are not similar; hence they tend to connect to molecules that share different properties in a way that can prove beneficial to heal defects in the longer term. The evolution of the largest cluster, represented as a graph, during assembly in water, is shown in **Figure 3**. The relevant evolution during assembly in DMF is available in the SI, section II.

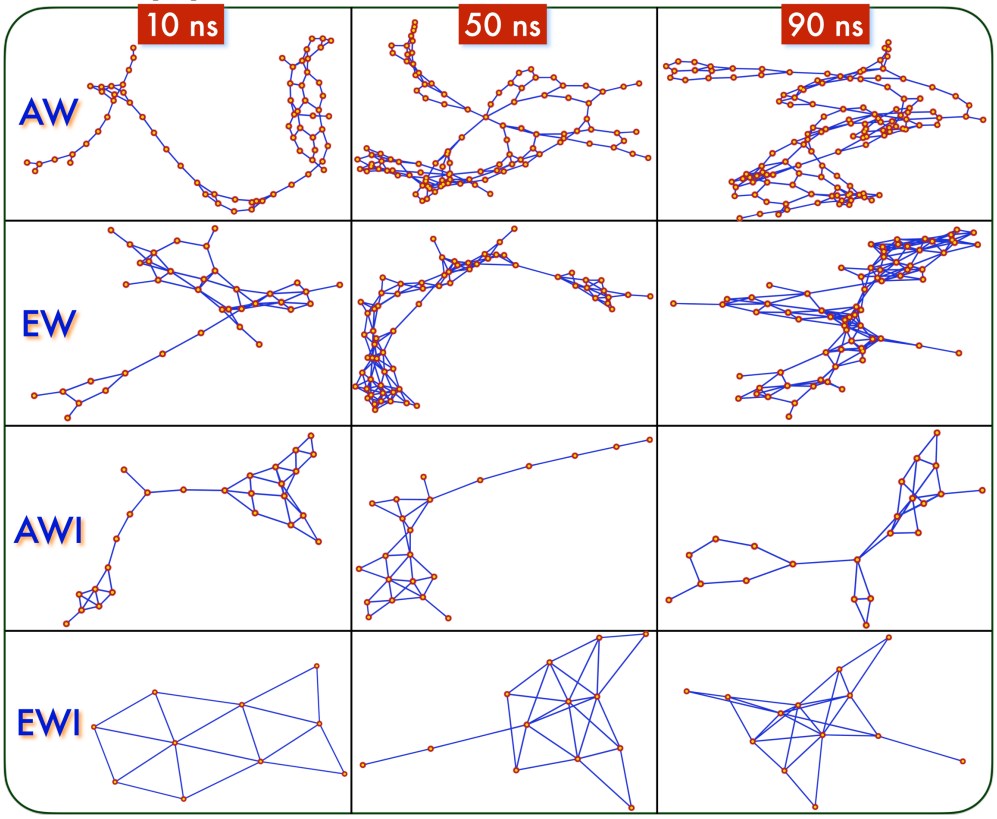


Figure 3. Graph representation of the largest cluster formed at 10, 50 and 90 ns of the production simulations in water. Code for abbreviations follows, AW: "Purely MLA in pure water", EW: "Equiprobable MLA/B/C in pure water", AWI: "Purely MLA in water with ions", EWI: "Equiprobable MLA/B/C in water with ions", AD: "Purely MLA in pure DMF", ED: "Equiprobable MLA/B/C in pure DMF", ADI: "Purely MLA in DMF with ions", EDI: "Equiprobable MLA/B/C in DMF with ions". Graph color code is the same as in **Figure 2**.

Estimation of rates. The rate of cluster formation during self-assembly (see **Figure 4**) is calculated for all simulations performed. Except from assembly in DMF with ions, samples with purely MLA half-SBUs invariably result in faster assembly rates than the equiprobable distribution. In all cases, ions tend to slow down self-assembly. However, ions can help heal defects. Pure DMF results in faster assembly for both solute compositions. The fastest assembly corresponds to the case of purely MLA half-SBUs in DMF. Finally, MLA forms clusters quicker in water with ions than in DMF with ions, while the opposite occurs for the equiprobable distribution of half-SBUs. The resulting values are provided in the SI, section VI.

Fractal growth process. We calculate the radii of gyration of clusters (see **Figure 4**) formed and growing in time during the simulation. We then relate these values with their corresponding size expressed in terms of number of half-SBUs. In this manner, we can calculate the fractal dimension associated with the growth process (see **Figure 4**). In pure water, we observe that in the "purely MLA" system evolves by forming two-dimensional structures as the corresponding fractal dimension is close to 2. Introduction of MLB and MLC leads to a slightly higher fractal dimension, while ions result in an overall decrease of the fractal dimension. On the other hand, in DMF, similar fractal dimension values are obtained for both solute compositions. However, these are appreciably lower than the ones in pure water; hence DMF decreases the fractal dimension related with self-assembly. Spectator ions in DMF further induce a slight decrease in the fractal dimension related with growth.

Diffraction pattern. In this section, we assess structural similarity based on the diffraction patterns of clusters formed during the simulation and that of MIL-101(Cr)[63] (reference structure). We performed this analysis both on clusters formed at the end of our simulations as well as the 5 largest clusters emerging every nanosecond during the whole simulation trajectory. Clusters formed out of MLA units are more similar to the reference based on cosine similarity. This more pronounced in the case of pure water. Consequently, MLA promotes the crystallinity of clusters formed during assembly in pure water. This analysis is provided in detail in the SI, section VII.

Principal Component Analysis. Finally, we calculate the principal components on which the descriptors of the nucleaton process computed from simulation can be projected. This analysis enables us to identify similar cases (e. g. purely MLA in water and DMF) as well as "unique" ones (equiprobable distributions in water with or without ions). This successful dimensionality reduction resulting from different sets of data agrees with the qualitative picture that we get from describing molecular behavior in the trajectories of different simulations. Results are available in **Figure 4**. At last, we see that the 1st principal component is highly correlated (more than 89%) with 2/3 of the data. This means that it can be used as a collective variable that can explain the early stages of growth as it incorporates most of the information gained from various quantities. At last, correlation coefficients involving this data and PCA of the observables at the very early stages of assembly are available in the SI, section VIII.

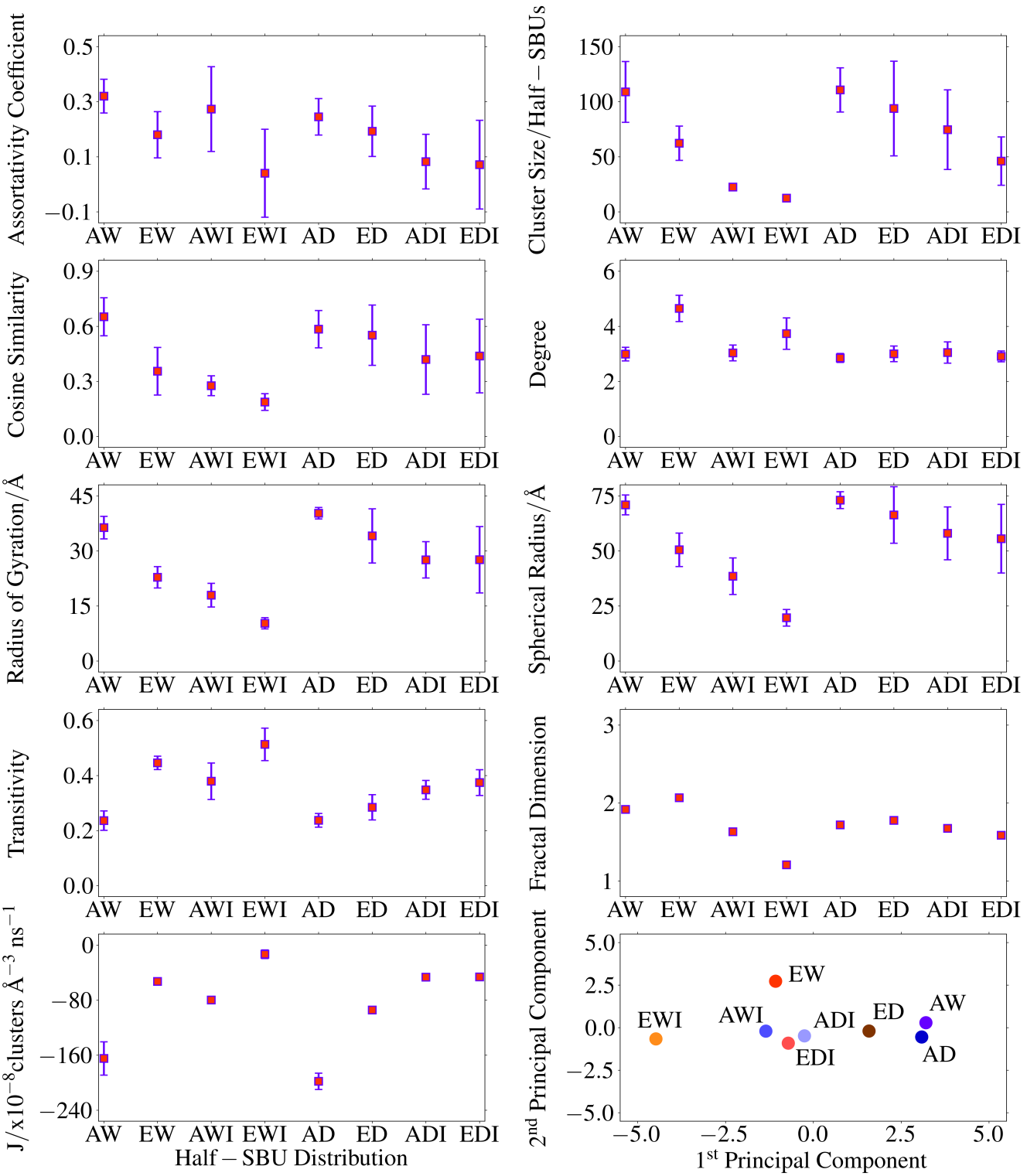


Figure 4. Principal Component Analysis (top center panel) based on nine dimensions. Dimensions are average quantities (squares) along with the corresponding standard deviations (error-bars). These are calculated over the production simulation trajectory. Cluster size, degree, transitivity, spherical radius, radius of gyration, assortativity coefficient, and cosine similarity refer to the largest cluster. Code for abbreviations is the same as in **Figure 3**.

CONCLUSIONS AND OUTLOOK

The early stages of MOF self-assembly are investigated through molecular simulation that allows us to identify key interactions during nucleation at the molecular level. We assess the effects of building unit distribution, solvent, and spectator ions on MIL-101 self-assembly by 100 ns-long simulations of more than 780,000 atoms. Clusters of building units are then characterized as an evolving graph. Our results corroborate previous studies, where the solution composition considerably influences the dynamical rearrangement of building units.[41, 64] Consequently, our conclusions confirm these phenomena at large length scales where the complexity is significantly increased.

This increased complexity can be represented through graph theory which we show can be used to monitor and characterize growth at the early stages by projecting trajectories of atomic positions on lower-dimension graphs. The graphs are used to interpret how the connectivity, size and morphology of clusters evolve during assembly. As a result, this work sets the base for further analysis of nucleation using data science to evaluate dynamics as the propagation of a coarse-grained graph model.

We conclude that a couple of molecular descriptors suffice to account for 90% the variance of the dataset. Consequently, we can successfully identify nucleation descriptors that capture the evolution of both local and extended features. One principal component is strongly correlated with the fractal dimension related with growth that is also a proxy for the average degree of nodes in the graph. As such, this descriptor allows us to track the evolution of the local structural environment of the half-SBU, a variable that is accessible through X-Ray absorption and scattering methods.[20, 72, 73] The other principal component is correlated with the cosine similarity that can be calculated from powder diffraction patterns, and measures long-range order. The latter can be monitored by time-resolved diffraction.[20, 28-32, 63, 72, 73] Ultimately, these descriptors can also form a basis for collective variables to simulate nucleation via enhanced sampling.[41, 74] This opens up the exciting possibility to monitor the evolution of these graphs both experimentally and computationally.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Supplementary information (SI) linked with the main article (PDF)

AUTHOR INFORMATION

Corresponding Author

*vanda.glezakou@pnnl.gov
*m.salvalaglio@ucl.ac.uk

Notes

The authors declare no competing financial interest.

ORCID

Loukas Kollias: 0000-0002-2114-1238
Roger Rousseau: 0000-0003-1947-0478
Vassiliki-Alexandra Glezakou: 0000-0001-6028-7021
Matteo Salvalaglio: 0000-0003-3371-2090

ACKNOWLEDGMENT

The work described in this publication was performed by Pacific Northwest National Laboratory, which is operated by Battelle for the United States Department of Energy under Contract DE-AC05-76RL0180. V.-A. G. and L. K. gratefully acknowledge support from by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division, project 72353 (Interfacial Structure and Dynamics in Ion Separations). The authors acknowledge the use of the UCL Grace, Myriad and Kathleen High Performance Computing Facility (Grace@UCL, Myriad@UCL and Kathleen@UCL), and associated support services, in the completion of this work. We are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science of the U.S. Department of Energy under Contract No. DE-SUPPLEMENTARY MATERIAL AC02-05CH11231.

REFERENCES

1. Qazvini, O. T.; Babarao, R.; Telfer, S. G., Selective capture of carbon dioxide from hydrocarbons using a metal-organic framework. *Nat. Commun.* **2021,** *12* (1).
2. Asgari, M.; Jawahery, S.; Bloch, E. D.; Hudson, M. R.; Flacau, R.; Vlaisavljevich, B.; Long, J. R.; Brown, C. M.; Queen, W. L., An experimental and computational study of CO2 adsorption in the sodalite-type M-BTT (M = Cr, Mn, Fe, Cu) metal–organic frameworks featuring open metal sites. *Chem. Sci.* **2018,** *9* (20), 4579-4588.
3. Cui, X.; Niu, Z.; Shan, C.; Yang, L.; Hu, J.; Wang, Q.; Lan, P. C.; Li, Y.; Wojtas, L.; Ma, S.; Xing, H., Efficient separation of xylene isomers by a guest-responsive metal–organic framework with rotational anionic sites. *Nat. Commun.* **2020,** *11* (1).
4. Hanikel, N.; Prévot, M. S.; Yaghi, O. M., MOF water harvesters. *Nat. Nanotechnol.* **2020,** *15* (5), 348-355.
5. Sheberla, D.; Bachman, J. C.; Elias, J. S.; Sun, C.-J.; Shao-Horn, Y.; Dincă, M., Conductive MOF electrodes for stable supercapacitors with high areal capacitance. *Nat. Mater.* **2016,** *16* (2), 220-224.
6. Horcajada, P.; Serre, C.; Maurin, G.; Ramsahye, N. A.; Balas, F.; Vallet-Regí, M. a.; Sebban, M.; Taulelle, F.; Férey, G. r., Flexible Porous Metal-Organic Frameworks for a Controlled Drug Delivery. *J. Am. Chem. Soc.* **2008,** *130* (21), 6774-6780.
7. Kirchon, A.; Feng, L.; Drake, H. F.; Joseph, E. A.; Zhou, H.-C., From fundamentals to applications: a toolbox for robust and multifunctional MOF materials. *Chem. Soc. Rev.* **2018,** *47* (23), 8611-8638.
8. Yuan, S.; Feng, L.; Wang, K.; Pang, J.; Bosch, M.; Lollar, C.; Sun, Y.; Qin, J.; Yang, X.; Zhang, P.; Wang, Q.; Zou, L.; Zhang, Y.; Zhang, L.; Fang, Y.; Li, J.; Zhou, H.-C., Stable Metal-Organic Frameworks: Design, Synthesis, and Applications. *Adv. Mater.* **2018,** *30* (37).

9. Cliffe, M. J.; Hill, J. A.; Murray, C. A.; Coudert, F.-X.; Goodwin, A. L., Defect-dependent colossal negative thermal expansion in UiO-66(Hf) metal–organic framework. *Phys. Chem. Chem. Phys.* **2015,** *17* (17), 11586-11592.

10. Dissegna, S.; Epp, K.; Heinz, W. R.; Kieslich, G.; Fischer, R. A., Defective Metal-Organic Frameworks. *Adv. Mater.* **2018,** *30* (37).

11. Taddei, M., When defects turn into virtues: The curious case of zirconium-based metal-organic frameworks. *Coord. Chem. Rev.* **2017,** *343*, 1-24.

12. Lee, T.; Chang, Y. H.; Lee, H. L., Crystallization process development of metal–organic frameworks by linking secondary building units, lattice nucleation and luminescence: insight into reproducibility. *CrystEngComm* **2017,** *19* (3), 426-441.

13. Férey, G.; Haouas, M.; Loiseau, T.; Taulelle, F., Nanoporous Solids: How Do They Form? An In Situ Approach. *Chem. Mater.* **2013,** *26* (1), 299-309.

14. Haouas, M.; Volkringer, C.; Loiseau, T.; Férey, G.; Taulelle, F., In Situ NMR, Ex Situ XRD and SEM Study of the Hydrothermal Crystallization of Nanoporous Aluminum Trimesates MIL-96, MIL-100, and MIL-110. *Chem. Mater.* **2012,** *24* (13), 2462-2471.

15. Haouas, M., Nuclear Magnetic Resonance Spectroscopy for In Situ Monitoring of Porous Materials Formation under Hydrothermal Conditions. *Materials* **2018,** *11* (8).

16. Xu, H.; Sommer, S.; Broge, N. L. N.; Gao, J.; Iversen, B. B., The Chemistry of Nucleation: In Situ Pair Distribution Function Analysis of Secondary Building Units During UiO-66 MOF Formation. *Chem. Eur. J.* **2019,** *25* (8), 2051-2058.

17. Howarth, A. J.; Peters, A. W.; Vermeulen, N. A.; Wang, T. C.; Hupp, J. T.; Farha, O. K., Best Practices for the Synthesis, Activation, and Characterization of Metal–Organic Frameworks. *Chem. Mater.* **2016,** *29* (1), 26-39.

18. Embrechts, H.; Kriesten, M.; Hoffmann, K.; Peukert, W.; Hartmann, M.; Distaso, M., Elucidation of the Formation Mechanism of Metal–Organic Frameworks via in-Situ Raman and FTIR Spectroscopy under Solvothermal Conditions. *J. Phys. Chem. C* **2018,** *122* (23), 12267-12278.

19. Liu, X.; Chee, S. W.; Raj, S.; Sawczyk, M.; Král, P.; Mirsaidov, U., Three-step nucleation of metal–organic framework nanocrystals. *Proc. Natl. Acad. Sci. U.S.A.* **2021,** *118* (10).

20. Van Vleet, M. J.; Weng, T.; Li, X.; Schmidt, J. R., In Situ, Time-Resolved, and Mechanistic Studies of Metal–Organic Framework Nucleation and Growth. *Chem. Rev.* **2018,** *118* (7), 3681-3721.

21. Yoneya, M.; Tsuzuki, S.; Aoyagi, M., Simulation of metal–organic framework self-assembly. *Phys. Chem. Chem. Phys.* **2015,** *17* (14), 8649-8652.

22. Biswal, D.; Kusalik, P. G., Probing Molecular Mechanisms of Self-Assembly in Metal–Organic Frameworks. *ACS Nano* **2016,** *11* (1), 258-268.

23. Wells, S. A.; Cessford, N. F.; Seaton, N. A.; Düren, T., Early stages of phase selection in MOF formation observed in molecular Monte Carlo simulations. *RSC Adv.* **2019,** *9* (25), 14382-14390.

24. Colón, Y. J.; Guo, A. Z.; Antony, L. W.; Hoffmann, K. Q.; de Pablo, J. J., Free energy of metal-organic framework self-assembly. *J. Chem. Phys.* **2019,** *150* (10).

25. Cantu, D. C.; McGrail, B. P.; Glezakou, V. A., Formation Mechanism of the Secondary Building Unit in a Chromium Terephthalate Metal-Organic Framework. *Chem. Mater.* **2014,** *26* (22), 6401-6409.

26. Nguyen, V.; Grünwald, M., Microscopic Origins of Poor Crystallinity in the Synthesis of Covalent Organic Framework COF-5. *J. Am. Chem. Soc.* **2018,** *140* (9), 3306-3311.

27. Maula, T. A.; Hatch, H. W.; Shen, V. K.; Rangarajan, S.; Mittal, J., Designing molecular building blocks for the self-assembly of complex porous networks. *Mol. Syst. Des. Eng.* **2019,** *4* (3), 644-653.

28. Ahn, S.; Jung, S. Y.; Kim, S.; Lee, S. J., Structure-dependent light-responsiveness of chemically linked nanoparticle clusters. *RSC Adv.* **2013,** *3* (4), 1055-1060.

29. Wang, Z.; Liu, T.; Jiang, L.; Asif, M.; Qiu, X.; Yu, Y.; Xiao, F.; Liu, H., Assembling Metal–Organic Frameworks into the Fractal Scale for Sweat Sensing. *ACS Appl. Mater. Interfaces* **2019,** *11* (35), 32310-32319.

30. Zhang, Z.; Liu, Y.-W.; Tian, H.-R.; Li, X.-H.; Liu, S.-M.; Lu, Y.; Sun, Z.-X.; Liu, T.; Liu, S.-X., Polyoxometalate-Based Metal-Organic Framework Fractal Crystals. *Matter* **2020,** *2* (1), 250-260.

31. Maurer, C.; Baumgartner, B.; Pabisch, S.; Akbarzadeh, J.; Peterlik, H.; Schubert, U., Porous titanium and zirconium oxo carboxylates at the interface between sol–gel and metal–organic framework structures. *Dalton Trans.* **2014,** *43* (3), 950-957.

32. Tsao, C.-S.; Yu, M.-S.; Chung, T.-Y.; Wu, H.-C.; Wang, C.-Y.; Chang, K.-S.; Chen, H.-L., Characterization of Pore Structure in Metal–Organic Framework by Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **2007,** *129* (51), 15997-16004.

33. Goesten, M. G.; Stavitski, E.; Juan-Alcañiz, J.; Martiñez-Joaristi, A.; Petukhov, A. V.; Kapteijn, F.; Gascon, J., Small-angle X-ray scattering documents the growth of metal-organic frameworks. *Catal. Today* **2013,** *205*, 120-127.

34. Petrus, E.; Segado, M.; Bo, C., Nucleation mechanisms and speciation of metal oxide clusters. *Chem. Sci.* **2020,** *11* (32), 8448-8456.

35. Qu, H.; Huang, Z.; Dong, X.; Wang, X.; Tang, X.; Li, Z.; Gao, W.; Liu, H.; Huang, R.; Zhao, Z.; Zhang, H.; Yang, L.; Tian, Z.; Cao, X., Truncated Face-Rotating Polyhedra Constructed from Pentagonal Pentaphenylpyrrole through Graph Theory. *J. Am. Chem. Soc.* **2020,** *142* (38), 16223-16228.

36. Janet, J. P.; Kulik, H. J., Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017,** *121* (46), 8939-8954.

37. Anderson, R.; Gómez-Gualdrón, D. A., Increasing topological diversity during computational "synthesis" of porous crystals: how and why. *CrystEngComm* **2019,** *21* (10), 1653-1665.

38. Boyd, P. G.; Woo, T. K., A generalized method for constructing hypothetical nanoporous materials of any net topology from graph theory. *CrystEngComm* **2016,** *18* (21), 3777-3792.

39. Ohno, H.; Mukae, Y., Machine Learning Approach for Prediction and Search: Application to Methane Storage in a Metal–Organic Framework. *J. Phys. Chem. C* **2016,** *120* (42), 23963-23968.

40. Xie, T.; France-Lanord, A.; Wang, Y.; Shao-Horn, Y.; Grossman, J. C., Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nat. Commun.* **2019,** *10* (1).

41. Kollias, L.; Cantu, D. C.; Tubbs, M. A.; Rousseau, R.; Glezakou, V. A.; Salvalaglio, M., Molecular Level Understanding of the Free Energy Landscape in Early Stages of Metal-Organic Framework Nucleation. *J. Am. Chem. Soc.* **2019,** *141* (14), 6073-6081.

42. Hagberg, A. A.; Schult, D. A.; Swart, P. J. In *Exploring Network Structure, Dynamics, and Function using NetworkX*, Proceedings of the 7th Python in Science Conference, Varoquaux, G. e.; Vaught, T.; Millman, J., Eds. 2008; pp 11-15.

43. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential

Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.

44. Jorgensen, W. L.; Tirado-Rives, J., The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988,** *110* (6), 1657-66.

45. Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D., Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012,** *8* (1), 61-74.

46. van der Spoel, D.; van Maaren, P. J.; Caleman, C., GROMACS molecule & liquid database. *Bioinformatics* **2012,** *28* (5), 752-753.

47. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M., LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997,** *18* (12), 1463-1472.

48. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993,** *98* (12), 10089-10092.

49. Bussi, G.; Donadio, D.; Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007,** *126* (1), 014101.

50. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984,** *81* (8), 3684-3690.

51. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981,** *52* (12), 7182-7190.

52. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015,** *1-2*, 19-25.

53. Humphrey, W.; Dalke, A.; Schulten, K., VMD - Visual Molecular Dynamics. *J. Mol. Graph.* **1996,** *14*, 33-38.

54. Tribello, G. A.; Giberti, F.; Sosso, G. C.; Salvalaglio, M.; Parrinello, M., Analyzing and Driving Cluster Formation in Atomistic Simulations. *J. Chem. Theory Comput.* **2017,** *13* (3), 1317-1327.

55. Fixman, M., Radius of Gyration of Polymer Chains. *J. Chem. Phys.* **1962,** *36* (2), 306-310.

56. Enright, M. B.; Leitner, D. M., Mass fractal dimension and the compactness of proteins. *Phys. Rev. E* **2005,** *71* (1).

57. Russell, S. W.; Li, J.; Mayer, J. W., In situ observation of fractal growth during a-Si crystallization in a Cu3Si matrix. *J. Appl. Phys.* **1991,** *70* (9), 5153-5155.

58. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013,** *1* (1).

59. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013,** *68*, 314-319.

60. Hernández-Rivera, E.; Coleman, S. P.; Tschopp, M. A., Using Similarity Metrics to Quantify Differences in High-Throughput Data Sets: Application to X-ray Diffraction Patterns. *ACS Comb. Sci.* **2016,** *19* (1), 25-36.

61. Wan, K. X.; Vidavsky, I.; Gross, M. L., Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass. Spectrom.* **2002,** *13* (1), 85-8.

62. 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* (3), 261-272.

63. Ferey, G.; Mellot-Draznieks, C.; Serre, C.; Millange, F.; Dutour, J.; Surble, S.; Margiolaki, I., A chromium terephthalate-based solid with unusually large pore volumes and surface area. *Science* **2005,** *309* (5743), 2040-2.

64. Kollias, L.; Cantu, D. C.; Glezakou, V. A.; Rousseau, R.; Salvalaglio, M., On the Role of Enthalpic and Entropic Contributions to the Conformational Free Energy Landscape of MIL-101(Cr) Secondary Building Units. *Adv. Theory Simul.* **2020**.

65. Jones, C. L.; Hughes, C. E.; Yeung, H. H. M.; Paul, A.; Harris, K. D. M.; Easun, T. L., Exploiting in situ NMR to monitor the formation of a metal–organic framework. *Chem. Sci.* **2021,** *12* (4), 1486-1494.

66. Yuhara, D.; Barnes, B. C.; Suh, D.; Knott, B. C.; Beckham, G. T.; Yasuoka, K.; Wu, D. T.; Sum, A. K., Nucleation rate analysis of methane hydrate from molecular dynamics simulations. *Faraday Discuss.* **2015,** *179*, 463-474.

67. 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* (7825), 357-362.

68. Fornberg, B., Generation of finite difference formulas on arbitrarily spaced grids. *Math. Comput.* **1988,** *51* (184), 699-699.

69. Ringnér, M., What is principal component analysis? *Nat. Biotechnol.* **2008,** *26* (3), 303-304.

70. Jolliffe, I. T.; Cadima, J., Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A* **2016,** *374* (2065).

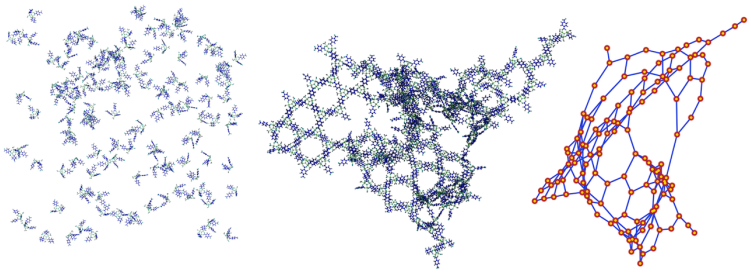
71. Lebedev, O. I.; Millange, F.; Serre, C.; Van Tendeloo, G.; Férey, G., First Direct Imaging of Giant Pores of the Metal−Organic Framework MIL-101. *Chem. Mater.* **2005,** *17* (26), 6525-6527.

72. Yang, F.; Hu, W.; Yang, C.; Patrick, M.; Cooksy, A. L.; Zhang, J.; Aguiar, J. A.; Fang, C.; Zhou, Y.; Meng, Y. S.; Huang, J.; Gu, J., Tuning Internal Strain in Metal–Organic Frameworks via Vapor Phase Infiltration for CO2 Reduction. *Angew. Chem. Int. Ed.* **2020,** *59* (11), 4572-4580.

73. Drisdell, W. S.; Poloni, R.; McDonald, T. M.; Long, J. R.; Smit, B.; Neaton, J. B.; Prendergast, D.; Kortright, J. B., Probing Adsorption Interactions in Metal–Organic Frameworks using X-ray Spectroscopy. *J. Am. Chem. Soc.* **2013,** *135* (48), 18183-18190.

74. Niu, H.; Piaggi, P. M.; Invernizzi, M.; Parrinello, M., Molecular dynamics simulations of liquid silica crystallization. *Proc. Natl. Acad. Sci. U.S.A.* **2018,** *115* (21), 5348-5352.

Graphic entry for the Table of Contents (TOC)



MOF half-SBUs **Molecular Cluster** **Graph Representation**