Eigencages: Learning a Latent Space of Porous Cage Molecules

Arni Sturluson, Melanie T. Huynh, Arthur H. P. York, Cory Simon

Submitted date: 13/11/2018 • Posted date: 14/11/2018
Licence: CC BY 4.0

Citation information: Sturluson, Arni; Huynh, Melanie T.; York, Arthur H. P.; Simon, Cory (2018): Eigencages: Learning a Latent Space of Porous Cage Molecules. ChemRxiv. Preprint.

Porous organic cage molecules harbor nano-sized cavities that can selectively adsorb gas molecules, lending them applications in separations and sensing. The geometry of the cavity strongly influences their adsorptive selectivity.

For comparing cages and predicting their adsorption properties, we embed/encode a set of 74 porous organic cage molecules into a low-dimensional, latent “cage space” on the basis of their intrinsic porosity. We first computationally scan each cage to generate a 3D image of its porosity. Leveraging the singular value decomposition, in an unsupervised manner, we then learn across all cages an approximate, lower-dimensional subspace in which the 3D porosity images lay. The “eigencages” are the set of orthogonal characteristic 3D porosity images that span this lower-dimensional subspace, ordered in terms of importance.

A latent representation/encoding of each cage follows from expressing it as a combination of the eigencages. We show that the learned encoding captures salient features of the cavities of porous cages and is predictive of properties of the cages that arise from cavity shape.
Eigencages: Learning a latent space of porous cage molecules

Arni Sturluson, Melanie T. Huynh, Arthur H. P. York, and Cory M. Simon*

Oregon State University, School of Chemical, Biological, and Environmental Engineering.
Corvallis, OR, USA.

E-mail: Cory.Simon@oregonstate.edu

Abstract

Porous organic cage molecules harbor nano-sized cavities that can selectively adsorb gas molecules, lending them applications in separations and sensing. The geometry of the cavity strongly influences their adsorptive selectivity. For comparing cages and predicting their adsorption properties, we embed/encode a set of 74 porous organic cage molecules into a low-dimensional, latent “cage space” on the basis of their intrinsic porosity. We first computationally scan each cage to generate a 3D image of its porosity. Leveraging the singular value decomposition, in an unsupervised manner, we then learn across all cages an approximate, lower-dimensional subspace in which the 3D porosity images lay. The “eigencages” are the set of orthogonal characteristic 3D porosity images that span this lower-dimensional subspace, ordered in terms of importance. A latent representation/encoding of each cage follows from expressing it as a combination of the eigencages. We show that the learned encoding captures salient features of the cavities of porous cages and is predictive of properties of the cages that arise from cavity shape.
Introduction

More than 10% of the world’s energy consumption is devoted to purifying chemical mixtures. The development of more energy-efficient processes to separate mixtures thus would significantly reduce carbon emissions and the cost of manufacturing goods. Gaseous mixtures in particular could be separated more energy-efficiently than by e.g. distillation instead via selective adsorption on a solid-state, porous material. Porous solids composed of porous organic cage molecules have auspiciously demonstrated the ability to separate gases in application to carbon dioxide capture from natural gas and flue gas of coal-fired power plants, xenon/krypton separations, and sulfur hexafluoride capture. Porous cage solids with high adsorptive selectivities may serve as vapor sensors as well.

Porous organic cages are molecules that harbor (i) a cavity that is intrinsic to their molecular structure and (ii) windows through which gas molecules can enter the cavity. Often, the cavity is large enough to accommodate only a single small gas molecule. Unlike metal- and covalent-organic frameworks, which are extended networks of molecular building blocks held together by directional coordination and covalent bonds, respectively, the assembly/packing of porous organic cage molecules to form a bulk porous cage solid is dictated by the geometry of the molecules and non-covalent/non-coordination interactions between them. On the order of hundreds of porous organic cages have been reported.

For deployment in molecular separations or vapor sensing, the size and shape of the cavities in a porous material can strongly influence the adsorptive selectivity. In a shape-selective molecular separation, the shape of the cavity or window in a porous material is such that it accommodates a subset of molecular species but excludes the remaining species through steric hindrance. Aside from geometric exclusion, the size and shape of the cavity influence the enthalpy of adsorption, e.g. by encompassing the adsorbed gas molecule with chemical moieties in close proximity with which to interact, and the entropy of adsorption, e.g. by minimizing the loss of rotational entropy in the cavity. Therefore, it is important to

---

1 The energy requirement for separating a mixture has a thermodynamic limit, of course.
mathematically characterize the geometry of the cavity/void space in nanoporous materials for predicting adsorption and comparing materials. Several methods to mathematically describe pores of materials include using the Voronoi decomposition, algebraic topology, radial distribution functions, and density of minima in the potential energy landscape. For extended networks, the MOFomics approach fits cylinders to the channels and describes the pore landscape as a graph of spheres (representing cages) connected to cylinders. Simple descriptors of cavities in porous cages such as the void and window diameters can be computed with pywindow.

In this work, our goal is to map porous organic cage molecules to a lower-dimensional latent space on the basis of their intrinsic porosity. i.e., we aim to develop an information-rich, low-dimensional vector representation—a fingerprint—of each porous organic cage that encodes the salient features of the size and shape of its cavity and windows. Such a fingerprint is useful for a few reasons. First, the latent representation would serve as a predictor of adsorption in a regression or classification model as in Refs.; see Le et al. for a review on quantitative structure-property modeling. Second, the latent representation lends a notion of similarity between two cages. Suppose a highly shape-selective porous cage molecule is composed of expensive or toxic precursors or suffers from instability. Within the latent cage space, we can then identify its nearest neighbors for alternative, existing cage molecules possessing the most similar cavity shapes, but composed of cheaper/safer precursors and showing higher stability. Finally, embedding porous cage molecules into a low-dimensional latent space and analyzing the clusters can shed light on the diversity of cavity shapes among the cages that have been synthesized.

Herein, we automatically learn a latent representation of the cavities of porous cages from a training dataset of 74 porous organic cage molecules. We achieve this by first computationally scanning the cage molecules to generate 3D images of their porosity, with the cavity in the center of the image. The pixels in the image represent a point in space and

*Latent* space refers to ‘hidden’ space: abstractly, it is a lower-dimensional subspace/ manifold—which is embedded in data space—in/on which the data congregates.
the binary pixel values represent whether the space is void or occupied by a cage atom. We postulate that these 3D porosity images, which belong to an enormous space, approximately lay in a much lower-dimensional subspace. Inspired by eigenfaces in facial recognition,\textsuperscript{38–40} we then employ the singular value decomposition to, in an unsupervised manner, identify a set of characteristic 3D porosity images—eigencages— that form an orthonormal basis for this approximate, lower-dimensional subspace in which the 3D porosity images lay. A low-dimensional latent representation follows by expressing each cage as a combination of the eigencages. Visualizing a 2D embedding of our latent cage space with t-SNE\textsuperscript{41} shows that the learned encoding captures salient features of the cavities of porous cages and is predictive of simulated xenon/krypton selectivity in isolated cage molecules.

**The porous organic cage molecule dataset**

Our training data set consists of the identity and spatial coordinates of the atoms comprising 74 porous organic cage molecules. A subset, 41 cages, were compiled by Miklitz et al.,\textsuperscript{16} who extracted cage molecules from crystal structures determined by X-ray diffraction studies and deposited in the Cambridge Structural Database\textsuperscript{42} by various research groups.\textsuperscript{5,43,44} The remaining cages are from a recent study that employed robots and computational predictions of cage topologies\textsuperscript{37} to synthesize 33 different cages. The atomic coordinates of the latter set of cage structures were obtained by molecular modeling, a subset of which were confirmed by X-ray diffraction studies.\textsuperscript{37} Fig. 1 illustrates the diversity of cage structures comprising the data set; Fig. S13.1 shows larger images of each cage. The spread of cavity and molecule diameters among the cages is shown in Fig. S1.1.

**Scanning the cages to generate 3D images of their porosity**

We describe here how we generate a raw 3D image of the porosity of a cage. These images can be conceptualized as a computerized tomography (CT) scan of a porous organic cage
Figure 1: The structures of all aligned 74 porous organic cage molecules comprising the training set in this study, ordered by cage molecule diameter computed from pywindow. Note the diversity of cavities.
Classifying a point as void space

We classify a given point in/around a cage as void space, as opposed to overlapping an atom of the cage, by computing the potential energy of a helium adsorbate at that point to determine if its interaction with the cage is dominantly repulsive. If and only if the potential energy is less than $k_B T$, with $k_B$ the Boltzmann constant and $T = 298$ K the temperature, the point is classified as void space. We model the energetics of the interaction of a helium atom with the atoms of the cage as pairwise additive and with 12-6 Lennard-Jones potentials, taking parameters from the Universal Force Field\textsuperscript{48} (geometric mixing rules, cutoff radius 14 Å). We use \texttt{PorousMaterials.jl v0.1.3}\textsuperscript{49} to compute the potential energies.

Alignment

The singular value decomposition is not equipped to learn rotational and translational invariance of cavity features in the 3D porosity images. Therefore, we must first consistently align the cages so that each pixel in the 3D porosity image corresponds to the same relative location for each cage. Alignment is also required for images of human faces for eigenfaces, so that e.g. the nose and eyes appear in the same region for each face.\textsuperscript{50} First, we translate each cage molecule so that its center of mass is at the origin. Second, we determine how to rotate each cage about its center of mass so that cages with similar cavity shapes are aligned.

As a basis for the rotational alignments, we first generate a set of points that fill and thus characterize the cavity of each cage. We successively insert helium adsorbates at random positions within a sphere circumscribing the cage molecule, rejecting the positions that overlap atoms of the cage and keeping the positions that were classified as void space; this set of points constitutes the porosity point cloud of a cage. We bias the insertions towards the center of the cage to emphasize the inner-cavity. See Sec. S2.1 for more details and Fig. S2.1 for example porosity point clouds.
Figure 2: Example 3D porosity images. (a-f) The molecular structure of a cage (name and experimental reference in subcaption) is shown along with a contour (0.5) of the 3D porosity image (orange). Bounding box shows $[-20 \, \text{Å}, 20 \, \text{Å}]^3$ dimension of the snapshot, consistent for all cages.
We use the rotational dynamics of the porosity point cloud as a first attempt to align the cages on the basis of their cavity shapes. We rotated each cage so that the principal axes of rotation of its porosity point cloud are aligned with the $x$, $y$, and $z$ Cartesian coordinate axes, with principal moments of inertia arranged in a non-increasing order. See Sec. S2.2. However, the cavity shapes of most cage molecules exhibit a degree of symmetry so as to render the computed principal axes of rotation sensitive to very small changes in the structure. i.e., the rotational dynamics of most porosity point clouds are approximately those of a sphere or disk, possessing nearly degenerate moments of inertia about the [therefore nearly arbitrary] principal axes of rotation. See Sec. S2.3 and Fig. S2.2 for examples of cages irrationally aligned on the basis of the rotational dynamics of their porosity point clouds. We give the principal axes of rotation authority to align a given cage only if the successive moments of inertia of the porosity point clouds about the principal axes of rotation differ by more than 1%, the case only for 18/74 cages (see Tab. S1). For the remaining cages whose porosity point clouds exhibited nearly degenerate principal axes of rotation, we resorted to using a point set registration algorithm, Coherent Point Drift.\textsuperscript{51}

We employed Coherent Point Drift\textsuperscript{51} to find the optimal rotation matrix to align the porosity point cloud of one cage, $\mathcal{Y}$ to the porosity point cloud of another cage, $\mathcal{X}$. Briefly, the Coherent Point Drift regards the points in $\mathcal{Y}$ as Gaussian mixture model centroids that generate the points in $\mathcal{X}$. The soft correspondences between points, i.e. the probability that a given point in $\mathcal{X}$ was generated by a Gaussian centered at a given point in $\mathcal{Y}$, are automatically inferred in Coherent Point Drift. The likelihood is maximized by the expectation-maximization (EM) algorithm: the E-step estimates probabilities of correspondence between the points given the current estimate for the rotation matrix and Gaussian variance; the M-step chooses the rotation matrix to minimize the distances between the points in $\mathcal{X}$ and transformed points in $\mathcal{Y}$, weighted by current estimates of the probabilities of correspondence, in addition to making the Gaussians as narrow as possible.

We then iteratively aligned each unaligned cage whose porosity point cloud possesses
nearly degenerate principal axes of rotation with an aligned cage. Let \( \mathcal{A} \) be the set of 18 cages aligned by the authority of the rotational dynamics of their porosity point clouds and \( \mathcal{U} \) be the set of remaining 56 cages not yet aligned. We use Coherent Point Drift to find which cage \( u \in \mathcal{U} \) aligns best (the largest likelihood) with a cage in \( \mathcal{A} \). We then align this cage \( u \) to the cage \( a \in \mathcal{A} \) with which it aligns best. We then remove \( u \) from the set of unaligned cages \( \mathcal{U} \) and add it to the set of aligned cages \( \mathcal{A} \) and repeat this process until \( \mathcal{U} \) is empty. See Tab. S2 for the directed graph of cage alignments.

In summary, we (i) translated each cage so its center of mass lays at the origin, (ii) generated a point cloud that fills the cavity of each cage, (iii) aligned each cage such that the principal axes of rotation of its porosity point cloud correspond with the Cartesian axes, then (iv) for cages whose porosity point clouds harbor insufficiently distinct principal moments of inertia, align them to another cage, on the basis of their porosity point clouds, using Coherent Point Drift.\(^{51}\)

**Generating the 3D porosity images**

For each aligned cage molecule, we overlay a regular, \( g \times g \times g \) grid of points (\( g = 50 \)) that span 40 Å in each dimension, centered at the center of mass of the cage. The dimension of the \([-20 \text{ Å}, 20 \text{ Å}]^3\) cubic grid of points was chosen as the smallest to encompass all atoms of all the cages. The 3D porosity image of a cage is then a 3D array; element \((i,j,k)\) is 0 if grid point \((i,j,k)\) is classified as void and 1 otherwise.

**Learning an approximate subspace of porosity images**

The raw representations of porous organic cage molecules as \( g \times g \times g \) 3D porosity images lie in a very high-dimensional space (\( g^3 = 125,000 \); flattening each image and viewing it as a vector). The main idea in this work is that the intrinsic cavities of porous cage molecules are not randomly distributed in this enormous space, but rather approximately lay in a much
lower-dimensional subspace of $\mathbb{R}^{g^3}$. As a revealing thought experiment, consider generating a random 3D porosity image by choosing each pixel as 0 or 1 randomly; it is extremely likely that this image will not resemble any known cage molecule. That is, the effective dimension of 3D porosity images is much lower than $g^3 = 125,000$.

We now leverage the singular value decomposition\textsuperscript{40,52,53} to learn from the set of 3D porosity images of the cages in Fig. 1 an approximate, lower-dimensional subspace in which the 3D porosity images of porous cage molecules lay. The \textit{eigencages} are a set of orthonormal vectors that span this lower-dimensional subspace; they are ordered in terms of which directions, in the space of all 3D porosity images, account for the most variance among the 3D porosity images of the cages in Fig. 1. Expressing a 3D porosity image as a combination of the eigencages then lends a latent representation of the cage.

**The data matrix, $A$**

We encapsulate all $c = 74$ 3D porosity images of porous cages into a data matrix $A$. We first flatten the $g \times g \times g$ ($g = 50$) images into a set of \textit{raw} vector representations \{$c_1, c_2, \ldots, c_c$\}, all of which lie in the enormous space $\mathbb{R}^{g^3}$. We compute the average 3D porosity image as $\bar{c} = \frac{1}{c} \sum_{i=1}^{c} c_i$ (visualized later in Fig. 5a). Now let $a_i := c_i - \bar{c}$ be the difference between the porosity image of cage $i$ and the average porosity image. The $c \times g^3$ data matrix $A$ is then defined by assigning its $i$th row to be $a_i^T$. As there are many fewer cages than pixels in the 3D porosity images ($c << g^3$), the data matrix $A$ is very wide (see Fig. S3.1).

**The singular value decomposition (SVD)**

The singular value decomposition (SVD)\textsuperscript{40,52,53} enjoys use in genomics,\textsuperscript{54} recommender systems,\textsuperscript{55} and image processing.\textsuperscript{40} We provide a pedagogical overview of the SVD for the reader’s convenience. We then reason that the SVD of our data matrix $A$ identifies an approximate lower dimensional subspace in which the 3D porosity images lay and a latent space of the cavities of the porous cages.
The matrix decomposition

The singular value decomposition (SVD) of the data matrix $A \in \mathbb{R}^{c \times g^3}$ ($g^3 > c$) is:

$$ A = U \Sigma V^T, \quad (1) $$

where $U$ is a $c \times c$ orthogonal matrix, $\Sigma$ is an $c \times g^3$ diagonal matrix with the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_c \geq 0$ of $A$ arranged down the diagonal, and $V$ is a $g^3 \times g^3$ orthogonal matrix. The columns of $U = [u_1, u_2, \ldots, u_c]$ form an orthonormal basis for $\mathbb{R}^c$ and are the left singular vectors of $A$; the columns of $V = [v_1, v_2, \ldots, v_{g^3}]$ form an orthonormal basis for $\mathbb{R}^{g^3}$ and are the right singular vectors of $A$. These left and right singular vectors are eigenvectors of $AA^T$ and $A^TA$, respectively; the non-zero singular values are the square roots of the [shared] non-zero eigenvalues of $AA^T$ and $A^TA$. Given that the latter matrix is proportional to the sample covariance matrix, we are essentially conducting principal component analysis, identifying the most significant directions of variance among the 3D porosity images.

Given that the matrix $A$ is rank $r$ and $\Sigma$ in eqn. 1 has many columns of zeros because our matrix is wide (see Fig. 3), we can write eqn. 1 in a reduced form:

$$ A = U_r \Sigma_r V_{r'}^T, \quad (2) $$

Figure 3: The full singular value decomposition of $A$ in eqn. 1. Relevant rows/columns are colored and labeled.
where now $\Sigma_r$ is a diagonal $r \times r$ matrix that contains only the $r$ non-zero singular values of $A$ down its diagonal in a non-increasing order, $U_r = [u_1 \ u_2 \ \cdots \ u_r]$ is a $c \times r$ matrix, and $V_r = [v_1 \ v_2 \ \cdots \ v_r]$ is a $g^3 \times r$ matrix. Writing eqn. 2 using an outer product expansion expresses $A$ as a sum of rank-one matrices:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T. \quad (3)$$

The singular values appear in eqn. 3 as weights on the rank-one matrices, formed by the outer product of two unit vectors, used to construct the matrix $A$. This emphasizes that the singular vectors are ordered in terms of significance.

We numerically compute the singular value decomposition using the $\text{svd}$ function in Julia.\textsuperscript{56}

**A geometric view of SVD**

A useful geometric view of the SVD follows by considering how $A$ maps a unit hypersphere in $\mathbb{R}^{g^3}$ into $\mathbb{R}^c$. We can express any point $x_s \in \mathbb{R}^{g^3}$ on the unit hypersphere as a linear combination of the right singular vectors:

$$x_s = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_{g^3} v_{g^3}, \quad (4)$$

such that $\sum_{i=1}^{g^3} \alpha_i^2 = 1$ to enforce $||x_s|| = 1$. Upon multiplication by $A$, the point $x_s$ is transformed to a new vector:

$$Ax_s = \alpha_1 \sigma_1 u_1 + \alpha_2 \sigma_2 u_2 + \cdots + \alpha_r \sigma_r u_r. \quad (5)$$

This follows from $Av_i = \sigma_i u_i$ for $i = 1, 2, ..., r$ and $Av_i = 0$ for $i = r + 1, ..., g^3$. Eqn. 5 describes an $r$-dimensional ellipsoid lying in $\mathbb{R}^c$, whose principal semi-axes are in the direction of $u_i$ with lengths of $\sigma_i$ for $i = 1, ..., r$. Therefore, multiplication by $A$ deforms the unit
hypersphere in \( \mathbb{R}^{g^3} \) by first collapsing \( g^3 - r \) dimensions, then stretching/compressing it along the remaining \( r \) dimensions, then rotating it, resulting in an \( r \)-dimensional ellipsoid that lays in \( \mathbb{R}^r \). The SVD recovers the directions of the principal semi-axes of this ellipsoid, \( \{u_i\} \), and their lengths, \( \{\sigma_i\} \), as well as each vector \( v_i \) that is mapped to \( \sigma_i u_i \). The vectors \( \{u_i\} \) are orthonormal and, remarkably, so are the set of vectors \( \{v_i\} \).

**Low-rank-\( \nu \) approximation \( A_\nu \) to the data matrix \( A \)**

Now, we employ the SVD to compress the data matrix \( A \) by finding a low-rank approximation. We define the “best” rank \( \nu < r \) approximation to \( A \), \( A_\nu \), as the one where \( ||A - A_\nu||_F \) is minimized, where \( || \cdot ||_F \) is the Frobenius norm. One can show that the optimal rank-\( \nu \) approximator is:

\[
A_\nu = \sum_{i=1}^{\nu} \sigma_i u_i v_i^T.
\]  

(6)

Comparing to eqn. 3, the optimal rank \( \nu \) approximator is obtained by setting the singular values \( \sigma_{\nu+1} = \sigma_{\nu+2} = \cdots = \sigma_r = 0 \). Aided by our geometric interpretation, we are approximating the linear transformation governed by \( A \) by collapsing the shortest principal axes of the ellipsoid in eqn. 5; justified intuitively, an ellipse e.g. in 2D is best-approximated by its longest principal axis. The relative error in the approximation is:

\[
\frac{||A - A_\nu||_F}{||A||_F} = \sqrt{\frac{\sum_{i=\nu+1}^{r} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2}}.
\]  

(7)

From a geometric standpoint, the relative error is related to the lengths of the principal semi-axes that we collapse in approximating the \( r \)-dimensional ellipsoid in eqn. 5 with a \( \nu \)-dimensional ellipsoid and how they compare to the longest principal semi-axes retained.
Interpreting the SVD for void space images of porous cage molecules

The data matrix $A$ encapsulates the 3D porosity images of all 74 porous cage molecules. We showed that we can approximate the data matrix $A$ with a lower-rank approximant $A_\nu$ in eqn. 6. The right singular vectors $\{v_i\}$ lie in the space of all 3D porosity images. As only the first $\nu$ right singular vectors appear in the approximant in eqn. 6, the best $\nu$-dimensional subspace of 3D porosity images is thus spanned by the orthonormal set of vectors $v_1, v_2, ..., v_\nu$. Analogous to eigenfaces, we declare this set of vectors, which are fictitious 3D porosity images [well, with $\bar{c}$ subtracted and normalized to have magnitude unity] discovered by SVD, the eigencages. Algebraically, eqn 6 approximates the flattened 3D porosity image of cage $k$ as:

$$c_k^T \approx \bar{c}^T + \sum_{i=1}^{\nu} \sigma_i u_i[k] v_i^T,$$

confirming cage $k$ is approximately a linear combination of the eigencages $v_1, v_2, ..., v_\nu$ with weights composed of the $k$th row of $U_\nu$ and the singular values $\sigma_1, \sigma_2, ..., \sigma_\nu$. In this linear combination, the singular value $\sigma_i$ appears as a weight to its corresponding eigencage $v_i$, indicating a hierarchy of importance of the eigencages and justifying discarding the singular vectors with smaller singular values in the approximant in eqn. 6. Instead of a $g^3$-dimensional 3D porosity image, each porous organic cage molecule can be represented by its composition of $v_1, v_2, ..., v_\nu$ given in eqn. 8. The latent representation of the 3D porosity image of cage $k$ is therefore row $k$ of $U_\nu \Sigma_\nu$. See Fig. 4.

We choose the dimension of the latent space $\nu$ as the smallest such that the relative error in eqn. 7 is less than 15%, leading to $\nu = 22$. See Fig. S5.1. As we have $c = 74$ cages, $\nu = c = 74$ would exactly reconstruct all cages in our cage data set. Thus, this compression to $\nu = 22$ dimensions is a 70% compression of the 3D porosity images of the cages while incurring only a 15% error in reconstructing the cages. The distribution of the singular values of $A$ is shown in Fig. S4.1.
To summarize, the singular value decomposition, in an unsupervised manner, learns:

• the best approximate \( \nu \)-dimensional subspace of \( \mathbb{R}^{g^3} \) in which the 3D images of the porosity of cage molecules lay. This subspace is spanned by the set of orthonormal right singular vectors \( \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\nu \), ranked in terms of importance, which we declare as eigencages.

• a \( \nu \)-dimensional latent space of porous organic cage molecules defined by the weights used to approximately construct a 3D porosity image from a linear combination of the eigencages. The \( \nu \)-dimensional latent representation of cage \( i \) is the \( i \)th row of \( \mathbf{U}_\nu \mathbf{\Sigma}_\nu \).

The eigencages

The eigencages– the rows of \( \mathbf{V}_\nu^T \) (see Fig. 4)– are an orthonormal basis for the approximate lower dimensional subspace in which all 3D porosity images lay and are ordered in terms of importance. The eigencages are the directions in 3D porosity image space that account for the most variance among the 3D porosity images in the dataset.\(^{53}\) We visualize the first (and most important) six eigencages in Fig. 5. As eqn. 8 illustrates, the eigencages express deviations from the average 3D porosity image \( \bar{c} \), whose contours are shown in Fig. 5a. Thus,
the eigencages possess contours at both positive and negative values. The periphery of the first eigencage $v_1$ in Fig. 5b exhibits radial symmetry; as Fig. S8.1 indicates, $v_1$ is used heavily to describe how the cavity diameter of a given cage differs from the average cage in Fig. 5a. The second and third eigencages appear to capture windows to the cavities and moieties that protrude from the core of the cage molecule. The fifth and sixth eigencages possess lobes likely describing windows but are difficult to reconcile with our intuition of how to describe the shape of a cage molecule, highlighting that a human-engineered feature of porous organic cages is unlikely to be optimal in compressing the information about a cage cavity into a low-dimensional representation.

**Reconstructing a cage from eigencages**

From the view of SVD, each 3D porosity image is constructed from a linear combination of the eigencages (see eqn. 8). We show here that, remarkably, even the six eigencages in Fig. 5 can be enough to visually reproduce the structure of the cavity in a porous organic cage molecule. As an example, in Fig. 6 we show the approximated reconstruction of the 3D porosity image of porous organic cage molecule $B_{25}$ using different numbers of eigencages. Mathematically, Fig. 6 shows the approximant given by eqn. 8 with $\nu = 1, 2, ..., 9$. Expressing $B_{25}$ in terms of the first two eigencages is insufficient to capture its shape; for $\nu = 4$, the cavity and windows of $B_{25}$ are captured to some extent. Only until we reach $\nu = 6$ does the outer-contour of the reconstructed 3D void space image of $B_{25}$ well-approximate the shape of the molecule. The reconstructed $B_{25}$ 3D porosity image with $\nu = 9$ is approaching visual indistinguishably from the exact 3D void space image in Fig. 6a. This shows that e.g. a $\nu = 9$-dimensional latent representation of $B_{25}$ contains sufficient information to visually reproduce the shape of its cavity.
Figure 5: Visualizing the eigencages. (a) Contour surfaces of the average 3D porosity image. Values increase from light to dark green. (b-g) Contour surfaces of the first six eigencages. Blue: low (negative), Gray: intermediate, red: high (positive).
Figure 6: Reconstructing cage $\mathbf{B_{25}}$ with its latent representation. (a) Exact 3D void space image of $\mathbf{B_{25}}$. (b-j) Reconstructions using latent representations of varying dimensions $\nu$. These are contours (0.5) of eqn. 8 with varying $\nu$. (Views from an angle to clearly see the cavity.)
Visualizing & interpreting the latent space of porous cages

The latent representations of the cavities of the porous organic cage molecules describe how each cage is composed of the eigencages and are the rows of $\mathbf{U}_\nu \mathbf{\Sigma}_\nu$ (see Fig. 4). Because the latent space is of too high of a dimension to visualize directly ($\nu = 22$), we resort to t-Distributed Stochastic Neighbor Embedding (t-SNE)\textsuperscript{41,57} to embed our learned latent space of porous cages into a two-dimensional space for visualization. In contrast to simply visualizing the first two dimensions of our latent space—essentially choosing $\nu = 2$ in eqn. 6—t-SNE is a non-linear dimensionality reduction algorithm and hence complicated structure in the $\nu = 22$-dimensional space can be retained/captured in the 2D embedding. Moreover, while SVD prioritizes maintaining large distances in data space in the low-dimensional space, t-SNE is designed to preserve the local structure of the high-dimensional data in the low-dimensional embedding, but it also often captures global structure.\textsuperscript{41} A tunable parameter in t-SNE is the perplexity, a loose measure of how many neighbors each data point has.\textsuperscript{41,57} We find that our clusters of cages are fairly robust to our choice of the perplexity (5, in the range recommended in Ref.\textsuperscript{41}). Fig. 7 shows the resulting 2D embedding of the 3D porosity images of the porous cages; each cage appears as a point in this visualization of our latent cage space.

The cavities of cages within clusters in the learned latent space appear strikingly similar. We highly encourage readers to explore our interactive visualization of the latent space at \url{simonensemble.github.io/latent-cage-space}; upon hovering the mouse over a point in latent space, an image of the cage structure displays to facilitate interpreting the latent space. Fig. 7 highlights a few salient clusters. The remaining clusters are shown in Fig. S6.1. The clustering of the cages in latent space according to the shape of their cavities is consistent with our intuition. As this clustering is learned automatically, we can easily generalize to hundreds of thousands of porous cages instead of manually grouping them together.

We duly address that any set of invented features of a porous organic cage molecule stacked into a vector technically serves as a latent space of cages. For example, let the first
Figure 7: The latent space of cages $U_\nu \Sigma_\nu$ embedded into 2D by t-SNE.\textsuperscript{41,57} Salient clusters are highlighted. See Fig. S6.1 for the remaining clusters.
entry of vector $\mathbf{x} \in \mathbb{R}^2$ be the diameter of the largest included sphere in the cavity and the second entry be the molecular mass of the cage; the set of all $\mathbf{x}$ defines a 2D latent space. However, our goal is to define a meaningful latent space in that it is predictive of properties; i.e. neighboring materials in a good latent space will exhibit similar properties.\footnote{Akin to the famous George Box adage that all models are wrong, but some are useful, we modify his adage here to “All latent spaces are wrong, some are useful.” } Compared to human-engineered/invented features of cages, we expect that our discovered latent space defined in $U_\nu \Sigma_\nu$ efficiently encodes the salient features of the cavities in porous organic cages: (i) we obtained the latent representations by optimally compressing the 3D porosity images with SVD; (ii) as Fig. 6 shows, the latent representation of a cage can visually reproduce the shape of its cavity; (iii) as Fig. 7 shows, clusters in latent space coincide with our intuition of similar cages.

To more rigorously judge the utility of the latent space that we learned with SVD, we assess if regions of latent cage space are correlated with cage properties. First, we investigate if neighboring cages in latent space tend to exhibit similar simulated equilibrium xenon/krypton selectivities. We place each cage molecule in an empty simulation box and compute the Henry coefficient of xenon and krypton ($T = 298$ K), then subtract the Henry coefficient of helium to mimic an excess adsorption experiment, as in Patil et al.\cite{Patil2010} See Sec. S7 for details, discussion, and comparison to experimental Xe and Kr adsorption measurements in noria (NC2)\cite{Patil2010} and CC3.\cite{Patil2009} The points in Fig. 8 are colored according to the simulated Xe/Kr selectivity at infinite dilution. A subset of the cages, marked as ‘X’s in Fig. 8, possess windows that are too narrow for xenon to percolate into the cavity, according to a potential energy barrier criterion; see Sec. S11. Clearly, neighboring cages in the latent space are more likely to exhibit similar Xe/Kr selectivities than cages further apart. Despite atom type not being explicitly fed into SVD to learn the latent cage space, the information about the shape of the cavity encoded into the latent representation is predictive of the simulated Xe/Kr selectivity in the isolated cage molecule. As a word of caution, however, we explain in the Discussion that the simulated Xe/Kr selectivity of an isolated cage is not necessarily a
quality prediction of the simulated Xe/Kr selectivity in the porous cage solid where the cage molecules are packed together. Second, we investigate if the molecule and cavity diameter of the cages computed by pywindow are correlated with the location of a cage in latent space. Fig. S8.2 clearly shows that cages nearby in latent space have a strong tendency to possess similar molecule and cavity diameters. Note that MC6, the largest cage in the database, appears to be an outlier. Finally, Fig. S9.1 shows that our latent representations nicely cluster together cages with the same number of windows in which gas molecules can enter the cavity. Together, Figs. 8, S8.2, and S9.1 show that our latent representation of porous cage molecules is useful for predicting properties that are correlated with the shape of the cavity (as is the case for Xe/Kr selectivity) and thus lends a meaningful notion of similarity between two cage molecules.

A walk through latent cage space

Gómez-Bombarelli et al. encoded SMILES strings of molecules into a latent space using neural networks and generated novel molecules by walking through latent molecule space. van Deursen and Reymond coin this as “chemical space travel”. Similarly, we show in Fig 9 that we can interpolate between two given cages in the $\nu = 22$ dimensional latent cage space to see how one cage morphs into another. While an interesting exercise to help interpret the latent space, walking in our latent cage space may be of limited utility since it remains unclear of how to synthesize a cage with a given cavity shape.

Flexible cages occupy a region of latent space

Thus far, we viewed the cage molecules as rigid. But, often cages are flexible to some degree, and the shape of their cavity fluctuates, which influences adsorption. To explore the effects of flexibility/structural fluctuations of cages on their latent representations, we performed molecular dynamics (MD) simulations in the NVT ensemble ($T = 298$ K) on four empty, isolated cages (CC2, CC3, CC4 and CC5) using a polymer specific
Figure 8: The latent space of cages $\mathbf{U}_\nu \mathbf{\Sigma}_\nu$ embedded into 2D by t-SNE.\textsuperscript{41,57} The color of points represents the simulated Xe/Kr selectivity of an isolated cage molecule in an empty box at 298 K. Points nearby in the latent cage space are likely to exhibit similar Xe/Kr selectivities. Cages marked with ‘X’ have windows too narrow for xenon to percolate into the cavity.
consistent force field (PCFF)\textsuperscript{65–67} as implemented by \texttt{DL\_FIELD v4.3}\textsuperscript{68} and a cage specific force field (CSFF).\textsuperscript{69} See Sec. S10 for details. We gathered snapshots of each cage while they were fluctuating in the MD simulation, computed their 3D porosity images, then projected them onto a $\nu = 2$-dimensional latent space. Fig. S10.2 shows that flexible cages, on the basis of the MD snapshots, each occupy a region in latent space. The size of the region is determined by the degree of flexibility of the cage; for example, the larger cage CC5 explores a larger region of latent space than CC3. Moreover, the latent regions may overlap (as the case with CC3 and CC4), indicating that, a fraction of the time, the cages exhibit a very similar cavity shape. Analyzing the region in latent space explored by flexible cages could provide more information about fluctuations of cavity shape than simple descriptors such as pore diameter since our latent representation can capture more complex features. Still, considering a rigid cage– the average configuration of a cage– is valuable while requiring significantly less computational expense.

Conclusions and Discussion

The idea to exploit different affinities of gases for a surface to purify a gaseous mixture is more than one hundred years old.\textsuperscript{70} Since, advanced classes of materials harboring nano-sized pores such as porous cage solids\textsuperscript{5,14} have emerged and offer high adsorptive selectivities.\textsuperscript{6–11} Because the size and shape of the cavity intrinsic to a porous organic cage molecule can strongly
influence selectivity for its deployment in a molecular separation or sensing application; it is important to mathematically characterize the void spaces of porous organic cage molecules for predicting adsorption and comparing materials.

In this study, we computationally scanned the porosity of 74 porous organic cage molecules to generate three-dimensional images, with the cavity in the center of the image. The flattened image serves as a raw vector representation of the intrinsic porosity of a cage that lies in a very high-dimensional space. We postulated that the 3D porosity images effectively lie in a much lower dimensional subspace of this enormous space. Using the singular value decomposition, we learned in an unsupervised manner the effective lower-dimensional subspace in which the void spaces of porous organic cages sit, which is characterized by a set of eigencages. Expressing a cage as a linear combination of eigencages defined its latent representation, which lent a notion of similarity between two cages. We embedded the latent representations of the porous organic cage molecules into two-dimensional space to visualize the clusters of cages. We found that the clusters in the learned latent space coincide with our intuition of cages exhibiting similarly-shaped cavities and that cages can be visually reconstructed from their low-dimensional latent representation. Furthermore, cages nearby in latent space are more likely to exhibit similar simulated Xe/Kr selectivities, cavity diameters, and number of windows entering the cavity. Together, this shows that our $\nu = 22$-dimensional latent representation efficiently encodes the salient features of the cavities in the porous cages displayed in Fig. 1.

We host an interactive visualization of the latent cage space at simonensemble.github.io/latent-cage-space. If a given cage exhibits a high adsorptive selectivity for a gas separation or sensing application as a consequence of its cavity shape, one can search nearby in latent cage space for likely similar performers.

The degree of compression of the 74 3D porosity images by the singular value decomposition sheds some light on the diversity of cages that have been synthesized. A $\nu = 22$ dimensional latent representation—a 70% compression of the 3D porosity images—incurred
less than 15% relative error. Consequently, one might suggest that the cavities of the 74 cages in Fig. 1 are composed of approximately 20 orthogonal cavity “motifs” (eigencages). As only 12 of the at least 20 probable porous organic cage molecule topologies have been synthesized, our latent space representation will be useful in comparing and predicting the properties of the many porous organic cage molecules that are likely to emerge in the future.

At this juncture, we mention the limitations of this work. First, often cages are found to be flexible as opposed to rigid as we took the cages here. In reality, we showed that cages occupy a region in latent space, not a point. Second, as we considered only a single cage molecule in isolation, the latent cage space includes only the notion of intrinsic void space, as opposed to the extrinsic void space that can arise from how the cages pack together to form the bulk solid. In fact, assembly of porous cage molecules to form a bulk solid can be sensitive to small changes in the cage molecule. Depending on the outer surface chemistry and geometry of the molecule, the assembly/packing of porous organic cage molecules can be such that the intrinsic pores in the bulk solid are isolated! For this reason, the simulated Xe/Kr selectivities in the isolated cage molecules in Fig. 8 may not necessarily be a quality prediction of the selectivity in the bulk cage solid. Our analysis of potential energies of adsorption sites in/on the cage molecules in Fig. S12.1, however, suggests that the dominant adsorption site in cages exhibiting the highest simulated Xe/Kr selectivity tend to be internal to their cavity; hence false positives may be unlikely unless the internal sites are blocked. See Sec. S7.1 for more discussion. A third limitation is that, while we may aim for the singular value decomposition to learn distinguishing features of the cavities of the cages, the algorithm takes notice of the moieties protruding from the core in addition to the internal cavity. This is evident by e.g. the distance between C9 and C1 and RCC1a and RCC1c in Fig. 7. A method to incentivize a dimensionality reduction algorithm to pay more attention to the center of the cage may be warranted if the internal cavity is the most important feature to describe. That said, the encoding of the periphery of the molecule may be desirable for

\[^{†}\]We encouraged SVD to focus on the cavity by (i) centering the cavity in the image and (ii) aligning cages based on the porosity point cloud tracing out the shape of the inner-cavity.
e.g. predicting how cage molecules assemble to form the solid. The final limitation is that the singular value decomposition is sensitive to alignments and translations for inter-cage comparisons. One’s mathematical definition of “optimal alignment” (e.g. using rotational dynamics or Coherent Point Drift) may not coincide with one’s intuition of how the cages should be aligned. Our future work is to employ rotational-invariant algorithms that detect features of cavities regardless of their orientation.

When computationally scanning the cage molecule to generate the 3D porosity image, we binarized each pixel as void space or overlapping the cage structure using helium as a probe. This ensures that the 3D porosity image only contains information about the shape of the cage molecule and its cavity. Another option to likely lend a more predictive feature for a regression model of adsorption of a particular gas is to instead assign the pixel value to be the free energy of adsorption of that gas centered at that point. For this study, we elected to binarize the pixels to yield a more generalizable latent representation and avoid overfitting the latent representation to describe a particular gas.

One may question why we do not employ t-SNE directly to reduce the dimensionality of the 3D porosity images. The inventors of t-SNE themselves first used principal component analysis (directly related to SVD) to reduce the dimensionality of images of handwritten digits preceding t-SNE to filter out noise and to speed up the computation. Our reasons for employing SVD instead of t-SNE directly are (i) the mapping by SVD lends more interpretability through visualizing the eigencages (Fig. 5), (ii) unlike t-SNE, SVD provides an explicit mapping function that can be used to (a) project unseen data into the latent space as we did with our MD snapshots and (b) generate fictitious pore structures as we did between NC2 and DC1 (iii) t-SNE is stochastic, whereas SVD is deterministic.

To further the ideas in this study, we are working towards embedding the pore structures of extended network materials such as metal-organic frameworks (MOFs) into a low-dimensional latent space. MOFs are more tunable materials than porous cages, as exemplified by the tens of thousands of MOFs that have been reported in comparison to the
hundreds of porous cage solids; thus, MOFs exhibit a greater diversity in their void space architectures. However, MOFs present a more formidable challenge than porous organic cage molecules because their 3D porosity images are periodic with varying Bravais lattices. We are exploring how more advanced dimensionality reduction algorithms such as autoencoders with convolutional layers may fare at learning/detecting hierarchal features of pore structures with translational and rotational invariance.

Acknowledgement

C.M.S. and A.S. thank the School of Chemical, Biological, and Environmental Engineering at Oregon State University for start-up funds. M.T.H. thanks the Pete and Rosalie Johnson Internship Program at Oregon State University. A.H.P.Y. thanks the URSA Engage Program at Oregon State University. We thank the anonymous reviewers for helpful comments that led to improvements in our results and the text.

Supplementary Information

A Jupyter Notebook with Julia code to reproduce the data in this article is available on Github at github.com/SimonEnsemble/latent_cage_space; the 3D porosity images and Henry coefficients were calculated with our open source code PorousMaterials.jl v0.1.3. A Supporting Information document is also available.

References

(1) Sholl, D. S.; Lively, R. P. Seven chemical separations: to change the world: purifying mixtures without using heat would lower global energy use, emissions and pollution—and open up new routes to resources. Nature 2016, 532, 435–438.

(2) Brueske, S.; Kramer, C.; Fisher, A. Bandwidth Study on Energy Use and Potential
Energy Saving Opportunities in U.S. Petroleum Refining. *Energetics Incorporated, prepared for the US Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office 2015*,

(3) Bhown, A. S.; Freeman, B. C. Analysis and status of post-combustion carbon dioxide capture technologies. *Environmental science & Technology* 2011, 45, 8624–8632.

(4) Tozawa, T. et al. Porous organic cages. *Nature Materials* 2009, 8, 973.

(5) Hasell, T.; Cooper, A. I. Porous organic cages: soluble, modular and molecular pores. *Nature Reviews Materials* 2016, 1, 16053.

(6) 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. *Angewandte Chemie International Edition* 2011, 50, 1046–1051.

(7) Tian, J.; Thallapally, P. K.; Dalgarno, S. J.; McGrail, P. B.; Atwood, J. L. Amorphous molecular organic solids for gas adsorption. *Angewandte Chemie International Edition* 2009, 48, 5492–5495.

(8) Hong, S.; Rohman, M. R.; Jia, J.; Kim, Y.; Moon, D.; Kim, Y.; Ko, Y. H.; Lee, E.; Kim, K. Porphyrin boxes: rationally designed porous organic cages. *Angewandte Chemie International Edition* 2015, 54, 13241–13244.

(9) Chen, L. et al. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nature Materials* 2014, 13, 954.

(10) Patil, R. S.; Banerjee, D.; Simon, C. M.; Atwood, J. L.; Thallapally, P. K. Noria: A Highly Xe-Selective Nanoporous Organic Solid. *Chemistry–A European Journal* 2016, 22, 12618–12623.

(11) Hasell, T.; Miklitz, M.; Stephenson, A.; Little, M. A.; Chong, S. Y.; Clowes, R.; Chen, L.; Holden, D.; Tribello, G. A.; Jelfs, K. E.; Cooper, A. Porous organic cages for
sulfur hexafluoride separation. *Journal of the American Chemical Society* 2016, 138, 1653–1659.

(12) Brutschy, M.; Schneider, M. W.; Mastalerz, M.; Waldvogel, S. R. Porous organic cage compounds as highly potent affinity materials for sensing by quartz crystal microbalances. *Advanced Materials* 2012, 24, 6049–6052.

(13) Brutschy, M.; Schneider, M. W.; Mastalerz, M.; Waldvogel, S. R. Direct gravimetric sensing of GBL by a molecular recognition process in organic cage compounds. *Chemical Communications* 2013, 49, 8398–8400.

(14) Holst, J. R.; Trewin, A.; Cooper, A. I. Porous organic molecules. *Nature Chemistry* 2010, 2, 915.

(15) Cooper, A. I. Porous molecular solids and liquids. *ACS Central Science* 2017, 3, 544–553.

(16) 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. *The Journal of Physical Chemistry C* 2017, 121, 15211–15222.

(17) Furukawa, H.; Cordova, K. E.; O’Keeffe, M.; Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* 2013, 341, 1230444.

(18) Diercks, C. S.; Yaghi, O. M. The atom, the molecule, and the covalent organic framework. *Science* 2017, 355, eaal1585.

(19) McKeown, N. B. Nanoporous molecular crystals. *Journal of Materials Chemistry* 2010, 20, 10588–10597.

(20) Evans, J. D.; Huang, D. M.; Haranczyk, M.; Thornton, A. W.; Sumby, C. J.; Doonan, C. J. Computational identification of organic porous molecular crystals. *CrystEngComm* 2016, 18, 4133–4141.
(21) 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. *Nature Chemistry* 2013, *5*, 276–281.

(22) Zhu, W.; Kapteijn, F.; Moulijn, J. Shape selectivity in the adsorption of propane/propene on the all-silica DD3R. *Chemical Communications* 1999, *2453–2454.*

(23) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. *Journal of Chemical Theory and Computation* 2018, *14*, 4427–4437.

(24) Smit, B.; Maesen, T. L. Towards a molecular understanding of shape selectivity. *Nature* 2008, *451*, 671–678.

(25) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal–organic frameworks. *Chemical Science* 2012, *3*, 2217–2223.

(26) Simon, C. M.; Mercado, R.; Schnell, S. K.; Smit, B.; Haranczyk, M. What are the best materials to separate a xenon/krypton mixture? *Chemistry of Materials* 2015, *27*, 4459–4475.

(27) Denayer, J. F.; Ocakoglu, R. A.; Arik, I. C.; Kirschhock, C. E.; Martens, J. A.; Baron, G. V. Rotational entropy driven separation of alkane/isoalkane mixtures in zeolite cages. *Angewandte Chemie* 2005, *117*, 404–407.

(28) Pinheiro, M.; Martin, R. L.; Rycroft, C. H.; Jones, A.; Iglesia, E.; Haranczyk, M. Characterization and comparison of pore landscapes in crystalline porous materials. *Journal of Molecular Graphics and Modelling* 2013, *44*, 208–219.
(29) Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *Journal of Chemical Information and Modeling* 2011, 52, 308–318.

(30) Lee, Y.; Barthel, S. D.; Dłotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications* 2017, 8, 15396.

(31) Fernandez, M.; Trefiak, N. R.; Woo, T. K. Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity. *The Journal of Physical Chemistry C* 2013, 117, 14095–14105.

(32) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *The Journal of Chemical Physics* 2009, 130, 104504.

(33) First, E. L.; Floudas, C. A. MOFomics: Computational pore characterization of metal–organic frameworks. *Microporous and Mesoporous Materials* 2013, 165, 32–39.

(34) Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *Journal of Chemical Information and Modeling* 2018.

(35) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-Based Descriptors to Rapidly Predict Hydrogen Storage in Metal-Organic Frameworks. *Molecular Systems Design & Engineering* 2018.

(36) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical Reviews* 2012, 112, 2889–2919.

(37) Greenaway, R.; Santolini, V.; Bennison, M.; Alston, B.; Pugh, C.; Little, M.; Miklitz, M.; Eden-Rump, E.; Clowes, R.; Shakil, A.; Cuthbertson, H.; Armstrong, H.;
Briggs, M.; Jelfs, K.; Cooper, A. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nature Communications* 2018, 9, 2849.

(38) Sirovich, L.; Kirby, M. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A* 1987, 4, 519–524.

(39) Turk, M. A.; Pentland, A. P. Face recognition using eigenfaces. Computer Vision and Pattern Recognition, 1991. Proceedings CVPR’91., IEEE Computer Society Conference on. 1991; pp 586–591.

(40) Muller, N.; Magaia, L.; Herbst, B. M. Singular value decomposition, eigenfaces, and 3D reconstructions. *SIAM Review* 2004, 46, 518–545.

(41) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008, 9, 2579–2605.

(42) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B: Structural Science* 2002, 58, 380–388.

(43) Evans, J. D.; Sumby, C. J.; Doonan, C. J. Synthesis and applications of porous organic cages. *Chemistry Letters* 2015, 44, 582–588.

(44) Zhang, G.; Mastalerz, M. Organic cage compounds–from shape-persistency to function. *Chemical Society Reviews* 2014, 43, 1934–1947.

(45) Crini, G. A history of cyclodextrins. *Chemical Reviews* 2014, 114, 10940–10975.

(46) Horng, Y.-C.; Lin, T.-L.; Tu, C.-Y.; Sung, T.-J.; Hsieh, C.-C.; Hu, C.-H.; Lee, H. M.; Kuo, T.-S. Preparation of a Reversible Redox-Controlled Cage-Type Molecule Linked by Disulfide Bonds. *European Journal of Organic Chemistry* 2009, 2009, 1511–1514.
(47) Ding, H.; Yang, Y.; Li, B.; Pan, F.; Zhu, G.; Zeller, M.; Yuan, D.; Wang, C. Targeted synthesis of a large triazine-based [4+ 6] organic molecular cage: structure, porosity and gas separation. *Chemical Communications* **2015**, *51*, 1976–1979.

(48) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(49) PorousMaterials.jl. 2018; [http://dx.doi.org/10.5281/zenodo.1400839](http://dx.doi.org/10.5281/zenodo.1400839).

(50) Zhang, S.; Turk, M. Eigenfaces. *Scholarpedia* **2008**, *3*, 4244.

(51) Myronenko, A.; Song, X. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *32*, 2262–2275.

(52) Kalman, D. A singularly valuable decomposition: the SVD of a matrix. *The College Mathematics Journal* **1996**, *27*, 2–23.

(53) Strang, G. *Introduction to Linear Algebra*; Wellesley-Cambridge Press Wellesley, MA, 2016; Vol. 5.

(54) Alter, O.; Brown, P. O.; Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **2000**, *97*, 10101–10106.

(55) Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008; pp 426–434.

(56) Bezanson, J.; Karpinski, S.; Shah, V. B.; Edelman, A. Julia: A fast dynamic language for technical computing. *SIAM Review* **2017**, *59*, 65–98.

(57) Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill* **2016**, 34
(58) Box, G. E. Science and Statistics. *Journal of the American Statistical Association* **1976**, *71*, 791–799.

(59) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science* **2018**, *4*, 268–276.

(60) van Deursen, R.; Reymond, J.-L. Chemical space travel. *ChemMedChem: Chemistry Enabling Drug Discovery* **2007**, *2*, 636–640.

(61) Camp, J. S.; Sholl, D. S. Transition state theory methods to measure diffusion in flexible nanoporous materials: application to a porous organic cage crystal. *The Journal of Physical Chemistry C* **2016**, *120*, 1110–1120.

(62) Holden, D.; Jelfs, K. E.; Trewin, A.; Willock, D. J.; Haranczyk, M.; Cooper, A. I. Gas diffusion in a porous organic cage: Analysis of dynamic pore connectivity using molecular dynamics simulations. *The Journal of Physical Chemistry C* **2014**, *118*, 12734–12743.

(63) Holden, D.; Chong, S. Y.; Chen, L.; Jelfs, K. E.; Hasell, T.; Cooper, A. I. Understanding static, dynamic and cooperative porosity in molecular materials. *Chemical Science* **2016**, *7*, 4875–4879.

(64) Witman, M.; Ling, S.; Jawahery, S.; Boyd, P. G.; Haranczyk, M.; Slater, B.; Smit, B. The influence of intrinsic framework flexibility on adsorption in nanoporous materials. *Journal of the American Chemical Society* **2017**, *139*, 5547–5557.

(65) Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. An ab initio CFF93 all-atom force field for polycarbonates. *Journal of the American Chemical Society* **1994**, *116*, 2978–2987.
(66) Sun, H. Force field for computation of conformational energies, structures, and vibrational frequencies of aromatic polyesters. *Journal of Computational Chemistry* **1994**, *15*, 752–768.

(67) Sun, H. Ab initio calculations and force field development for computer simulation of polysilanes. *Macromolecules* **1995**, *28*, 701–712.

(68) Yong, C. W. Descriptions and Implementations of DL_F Notation: A Natural Chemical Expression System of Atom Types for Molecular Simulations. *Journal of Chemical Information and Modeling* **2016**, *56*, 1405–1409, PMID: 27455451.

(69) Holden, D.; Jelfs, K. E.; Cooper, A. I.; Trewin, A.; Willock, D. J. Bespoke Force Field for Simulating the Molecular Dynamics of Porous Organic Cages. *The Journal of Physical Chemistry C* **2012**, *116*, 16639–16651.

(70) Boyle, R. The Absorption of the Radioactive Emanations by Charcoal. *The Journal of Physical Chemistry* **1908**, *12*, 484–506.

(71) Santolini, V.; Miklitz, M.; Berardo, E.; Jelfs, K. E. Topological landscapes of porous organic cages. *Nanoscale* **2017**, *9*, 5280–5298.

(72) Hasell, T.; Culshaw, J. L.; Chong, S. Y.; Schmidtmann, M.; Little, M. A.; Jelfs, K. E.; Pyzer-Knapp, E. O.; Shepherd, H.; Adams, D. J.; Day, G. M.; Cooper, A. Controlling the crystallization of porous organic cages: molecular analogs of isoreticular frameworks using shape-specific directing solvents. *Journal of the American Chemical Society* **2014**, *136*, 1438–1448.

(73) Moghadam, P. Z.; Li, A.; Wiggin, S. B.; Tao, A.; Maloney, A. G.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D. Development of a Cambridge Structural Database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials* **2017**, *29*, 2618–2625.
(74) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507.

(75) Kavukcuoglu, K.; Sermanet, P.; Boureau, Y.-L.; Gregor, K.; Mathieu, M.; Cun, Y. L. Learning convolutional feature hierarchies for visual recognition. *Advances in Neural Information Processing Systems* **2010**, 1090–1098.

(76) Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; Fergus, R. Deconvolutional networks. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2010**, 2528–2535.
Graphical TOC Entry
Supporting Information for: Eigencages: Learning a latent space of porous cage molecules

Arni Sturluson, Melanie T. Huynh, Arthur H. P. York, and Cory M. Simon*

Oregon State University, School of Chemical, Biological, and Environmental Engineering.
Corvallis, OR, USA.

E-mail: Cory.Simon@oregonstate.edu

S1 Distribution of cage molecule and cavity diameters

Figure S1.1: The distribution of cage molecule diameters (red) and cage cavity diameters (blue) among the 74 cages analyzed in this work. These diameters were computed with pywindow.1
S2  Aligning the cages before scanning for void space

S2.1  Porosity point clouds

As a basis to determine the rotational alignments of the cages that will best align groups of cages with similarly-shaped cavities, we first generate a set of points that ‘fill’ the porosity of each cage. As examples, Fig. S2.1 shows example porosity point clouds for NC2 and A11. We generate the porosity point cloud by successively inserting helium adsorbates into a sphere circumscribing each cage molecule. The radius of the sphere is chosen as the cage molecule radius. We add the position of the randomly inserted helium adsorbate to a stack if and only if the point is classified as void space, rejecting the points that were not classified as void space (determined by the potential energy of a helium adsorbate probe there). This stack of points that lie in the void space of the cage and thus trace out the shape of its cavity is the porosity point cloud of a cage. We biased insertions towards the center of the cavity to emphasize the inner-cavities of the cages. For each insertion proposal, we generated a vector \( r = [N_x, N_y, N_z] \) where \( N_{x,y,z} \) are Gaussian distributed numbers with unit variance and zero mean and \( r \) is the radius of the cage molecule; points outside of the sphere circumscribing the cage are rejected. The factor of \( 1/5 \) is to concentrate the porosity point cloud in the cavity instead of the periphery. This is especially important for the rotational dynamics, since the points furthest from the center have a greater influence on the rotational dynamics. We used 100,000 points to describe the porosity point cloud for computing rotational dynamics and 5,000 points for applying the Coherent Point Drift algorithm since the latter is more computationally expensive.

S2.2  First attempt: alignment on the basis of rotational dynamics

We compute the \( 3 \times 3 \) symmetric moment of inertia matrix \( M \) of the porosity point cloud of each cage, defined relative to the center of mass of the cage, and find its eigendecomposition \( M = W \Lambda W^T \), with the eigenvalues arranged in a non-increasing order down the diagonal.
Figure S2.1: The porosity point clouds (5000 points each) generated to describe the cavity of NC2 (left) and A11 (right) are shown along with the cage structures.

matrix \( \Lambda \) and the orthonormal eigenvectors in the columns of \( W \). The eigenvectors of the matrix \( M \) are the principal axes of inertia, and the associated eigenvalues are the moments of inertia with respect to those axes, the principal moments of inertia. Then, we transform the coordinates of the centered cage by left-multiplying by the rotation matrix \( W^T \). The porosity point cloud of the rotated cage then possesses a diagonal moment of inertia matrix with the moments of inertia arranged in a non-increasing order down the diagonal. Next, we will show that for some cages, the rotational dynamics of the porosity point cloud do not provide a reasonable means to consistently align them. For only 18 of the cages did we use the principal axes of inertia to align them.

S2.3 Degenerate principal axes of inertia

The porosity point clouds of cage molecules that exhibit a degree of symmetry can possess degenerate moments and principal axes of inertia. As a pedagogical example, consider sulfur trioxide (SO\(_3\)), a trigonal planar molecule, shown in Fig. S2.2a. Its moment of inertia matrix \( M \) about the center of the molecule (the S atom), with Cartesian directions defined in
Fig. S2.2a and the sulfur atom at the origin, is:

\[
M = m_0 \ell_{S-O}^2 \begin{bmatrix}
  3/2 & 0 & 0 \\
  0 & 3/2 & 0 \\
  0 & 0 & 3
\end{bmatrix},
\]  

(S1)

where \( m_0 \) is the mass of an oxygen atom and \( \ell_{S-O} \) is the sulfur-oxygen bond length. The eigenvalues of the moment of inertia matrix \( M \) and thus principal moments of inertia are \( 3/2, 3/2, \) and \( 3 \) times \( m_0 \ell_{S-O}^2 \). The eigenvector (unique if unit norm enforced) associated with the largest eigenvalue, \( 3m_0 \ell_{S-O}^2 \), is \([0, 0, 1]\), indicating that the principal axis of inertia with the largest moment of inertia is pointing out of the page with respect to Fig. S2.2a. The remaining two principal axes of rotation must be orthogonal to \([0, 0, 1]\) and thus lay in the \( x-y \) plane. The moments of inertia about the remaining two principal axes of rotation, however, are equal, as the remaining eigenvalues \( 3/2, 3/2 \) times \( m_0 \ell_{S-O}^2 \) are degenerate. In fact, any vector laying in the \( x-y \) plane is an eigenvector of the matrix \( M \) in eqn. S1 associated with eigenvalue \( 3m_0 \ell_{S-O}^2/2 \). This mathematically indicates that the two remaining principal axes of rotation for \( \text{SO}_3 \) are not unique; any two sets of orthogonal vectors lying in the \( x-y \) plane will serve as principal axes of rotation of the \( \text{SO}_3 \).

As a consequence, the principal axes of rotation for a symmetric molecule such as \( \text{SO}_3 \) do not provide a means to uniquely align the molecule. Thus, e.g. to compare \( \text{SO}_3 \) to a molecule exhibiting the same trigonal planar symmetry, such as boron trifluoride \( \text{BF}_3 \), the principal axes of rotation cannot be used to align the two molecules along their bonds (aside from the principal axes of rotation associated with the largest moment of inertia, \([0, 0, 1]\) in Fig. S2.2a). With regards to its rotational dynamics, \( \text{SO}_3 \) behaves as a uniform disk.\(^2\)

The cavities of several cage molecules exhibit a high degree of symmetry, so as to behave akin to \( \text{SO}_3 \), where the principal axes of rotation are not unique and thus cannot be used to consistently align them. Interestingly, juxtapose cage \( \text{B6} \) in Fig. S2.2b with \( \text{SO}_3 \), which resembles a three-blade propeller\(^2\) with nearly degenerate principal axes of rotation in one
Figure S2.2: When the principal axes of rotation are not uniquely defined. (a) Sulfur trioxide (SO$_3$) is a trigonal planar molecule with degenerate principal axes of rotation in the $x$-$y$ plane. (b) B6 is a cage molecule whose rotational dynamics resemble SO$_3$ and thus exhibits nearly degenerate principal axes of rotation in one plane.

While we focused on the porosity point clouds to align the cages more on the basis of their cavity shapes, as opposed to focusing directly on the atoms of the cage, the idea is the same: the porosity point clouds of many cages harbor rotational dynamics akin to a sphere or disk, where the principal axes of rotation are degenerate and thus sensitive to small differences in the structure/cavity shape.

For example, the three moments of inertia of the porosity point clouds of CC1 and of CC3 differ by less than 1\%, indicating that the rotational dynamics of CC1 and CC3 are akin to a sphere; therefore, using the principal axes of rotation to align CC1 and CC3 (i) theoretically will not provide a unique manner in which to align them and (ii) will be extremely sensitive to small perturbations in the structure and small differences in the random points within the porosity point cloud. Note that CC3 and CC1 differ only in the vertices external to the cavity, but, as Fig. S2.3 shows, CC1 and CC3 are aligned irrationally on the basis of their computed principal axes of rotation due to the [near-]degeneracy of their axes.
of rotation. Another example that is highly related to the \( \text{SO}_3 \) example in Fig. S2.2a are \( \text{B6} \) and \( \text{B8} \), also shown in Fig. S2.3. Here, small differences in the three propeller-blade-like moieties that protrude from \( \text{B6} \) and \( \text{B8} \) as well as small differences in the random positions in the porosity point cloud result in significantly inconsistent alignments for comparing \( \text{B6} \) and \( \text{B8} \). \( \text{B6} \) and \( \text{B8} \) much resemble the \( \text{SO}_3 \) example in Fig. S2.2a, as two principal axes of rotation are nearly degenerate and thus the principal axes in that plane (the plane of the page in Fig. S13.1) are very sensitive to small perturbations in the structure.

Therefore, the principle axes of rotation are in many cases insufficient to determine an alignment of a porous cage molecule that is unique and robust to small differences. The set of cages whose porosity point clouds display successive principal moments of inertia that differ by greater than one percent are in Table S1. Only for these cages did we give the principal axes of rotation the authority to align them. For the remaining cages, whose rotational dynamics resemble those of a sphere or disk with practically arbitrary principal axes of rotation, we resorted to employing the Coherent Point Drift algorithm to pairwise align cages as described in the “Alignment” section in the main text. For the remaining cages not present in Table S1, Table S2 shows the result of which cages were aligned to which cages (in order of top to bottom) on the basis of their porosity point clouds, using the Coherent Point Drift algorithm. The Julia code for our Coherent Point Drift algorithm can be found at https://github.com/SimonEnsemble/CoherentPointDrift.jl.

| Cage | aligned to... |
|------|---------------|
| CC2  | CP1           |
| WC4  | CP3           |
| HC1  | CP4           |
| CD2  | CP3           |
| CC9  | CP1           |
| CC4  | CP3           |
| CC3  | CP3           |
RCC3b  CP3
RCC1d  CB5
CC10  CP1
CD3  CB7
A11  CD1
B11  CD1
CC1  CP4
RCC1c  CB5
B23  CD1
CC5  CC1
C11  CB7
C23  CD1
MC4  CC1
MC7  CD1
WC1  CP4
B15  CC2
MC5  CD2
C15  CC2
B24  CD3
MC2  HC1
C24  IC2
DC1  CP1
MC1  HC1
B18  RCC1d
B25  NC2
C18  CC4
C25  CD3
| Cage  | Alignment |
|-------|-----------|
| GC1   | RCC1d     |
| B26   | RCC1d     |
| IC1   | CC3       |
| C1b   | CB5       |
| MC6   | B24       |
| RCC1a | CP4       |
| B1    | CP4       |
| B2    | B1        |
| B9    | B2        |
| C8    | B2        |
| C2    | B1        |
| C9    | C2        |
| B8    | C2        |
| C1    | B1        |
| C6    | B2        |
| B6    | C8        |
| C5    | C2        |
| C4    | B8        |
| B4    | B8        |
| B5    | B8        |
| C21   | B6        |

Table S2: List of cages whose porosity point clouds harbored rotational dynamics that approximate those of a sphere or disk and thus are not authoritative for aligning them. This table shows which cage these cages were aligned to, in order from top to bottom, on the basis of their porosity point clouds and using the Coherent Point Drift algorithm.
Figure S2.3: Cases where the computed principal axes of rotation of the porosity point clouds do not provide reasonable alignments of cages. Shown are the porosity point clouds and structures of four cages. (a, b) The rotational dynamics of CC1 and CC3 are akin to a sphere; the three moments of inertia of the porosity point clouds for each CC1 and CC3 are computed to differ less than 1% from one another. Shown are the two cages aligned with the computed principal axes of rotation of their porosity point clouds. Despite the structures differing only in the their vertices at the periphery of the cage, the two cages are aligned inconsistently/irrationally. (c, d) Shown are cages B6 and B8 aligned with the principal axes of rotation of their porosity point clouds. B6 and B8 have two nearly degenerate axes of rotation in the plane of the page. As a result, despite only subtle differences in the three propeller-blade-like moieties that protrude from their cores and small differences in the random positions in the porosity point clouds, these two cages are aligned irrationally.
Table S1: List of cages aligned on the basis of the authoritative rotational dynamics of their porosity point clouds.

|       |
|-------|
| B13   |
| C13   |
| C20   |
| CB5   |
| CB6   |
| CB7   |
| CD1   |
| CP1   |
| CP3   |
| CP4   |
| CP5   |
| IC2   |
| MC3   |
| NC1   |
| NC2   |
| RCC3a |
| WC2   |
| WC3   |
Figure S3.1: An attempt at visualizing the structure of the $c \times g^3$ data matrix $A$. The color depicts the magnitude of the entry of the matrix. This is an “attempt” because $g^3 >> c$, so the data matrix $A$ is wide and difficult to visualize without a longer page. Zoom in to see some structure. We remark that $48334/125000$ of the columns (representing pixels in the 3D porosity images of the cages) are non-zeros.
S4  The singular values of $A$

Figure S4.1: The distribution of the set of singular values $\sigma_1, \sigma_2, ..., \sigma_{74}$ of $A$. 
S5  Relative error from approximating $A$ as $A_\nu$

Figure S5.1: The relative error in approximating the data matrix $A$ with a lower-rank-$\nu$ approximant $A_\nu$ given in eqn. 6. The formula for the relative error as a function of $\nu$, related to the singular values of $A$, is given in eqn. 7.
Figure S6.1: The latent space of cages $U_{\nu} \Sigma_{\nu}$ embedded into 2D by t-SNE. Salient clusters are highlighted. See Fig. 7 for the remaining clusters.
We describe more details here of the Henry coefficient calculations used to obtain the data in Fig. 8. We model the energetics of the interaction of a gas atom (Xe, Kr, He) with the atoms of the cage as pairwise additive and with 12-6 Lennard-Jones potentials. The $\epsilon_i$ and $\sigma_i$ parameters of the Lennard-Jones potentials for atom $i$-atom $i$ interactions are taken from the Universal Force Field $^5$ (cutoff radius 14 Å). Geometric mixing rules are applied for cross-interactions.

We placed an isolated cage molecule in an empty simulation box and took it as rigid. If $x \in \mathbb{R}^3$ is the position of a gas atom in the simulation box, the molecular model described above gives us the potential energy of the gas molecule, $U(x)$.

Consider the isolated porous cage molecule immersed in a bath of gas; the simulation box is drawn around the porous cage molecule. Henry’s law models the density of gas in the simulation box at thermodynamic equilibrium, $\rho$, as $\rho = KP$, with $K$ the Henry coefficient and $P$ the pressure of the gas. Henry’s law is valid only at low surface coverage.

The Henry coefficient of a gas in the simulation box including the isolated cage molecule is given as:

$$K = \frac{1}{|\Omega|} \int_{\Omega} e^{-\beta U(x)} dx,$$

(S2)

where $\beta = (k_B T)^{-1}$ is the thermodynamic beta ($T$ temperature, $k_B$ Boltzmann constant) and the integral is over the simulation box $\Omega$ (which has volume $|\Omega|$). Note that if $U(x) = 0$, Henry’s law recovers the ideal gas law $\rho = \beta P$, where $|\Omega|$ is the volume of the simulation box.

We computed the average in eqn. S2 via Monte Carlo integration, i.e. Widom particle insertions $^6$ using 1000 Monte Carlo insertions per volume ($\text{Å}^3$) of the simulation box. We used PorousMaterials.jl v0.1.1 $^7$ to compute the Henry coefficient.
S7.1 Comparing to experimental data

By considering only an isolated cage molecule in a simulation box, as opposed to many cage molecules packed together to form a solid, we account for only the influence of the intrinsic porosity to the cage molecule on the adsorption. This is an approximation, as we are neglecting the influence of the extrinsic porosity that arises from how the cage molecules pack together to form a bulk solid. When the cage molecules pack together, the adsorption sites on the exterior of the molecule or at the cage windows can be modified depending on how the cage molecules pack together. However, when the interior of the cage molecule is the dominant adsorption site in the bulk solid, this approach may be a viable method to predict the Henry coefficient of a bulk porous cage solid.

To evaluate our crude method of considering an isolated cage molecule, Fig. S7.1 compares experimentally measured xenon and krypton adsorption isotherms in noria\(^8\) (NC2) and CC3\(^9\) to the resulting Henry’s law with the Henry coefficient obtained from our simulations. Fig. S7.1 shows this method yields a very good prediction of experimental Xe and Kr Henry coefficients in noria\(^8\) (NC2), but underestimates the Henry coefficients in CC3.\(^9\) See Patil et. al\(^8\) for more discussion.

Figure S7.1: Comparison of experimental xenon and krypton adsorption isotherms at 298 K in noria\(^8\) and CC3\(^9\) to the simulated adsorption. Points show experimental data and the lines show Henry’s law with the Henry coefficient taken from our simulations. The simulated Henry coefficient of He was subtracted to account for the empty space in the simulation box.
S8  Relationship of latent representation and cage molecule and cavity diameters

Fig. S8.1 shows that the first component of the learned latent representation of a cage is strongly correlated with its cavity diameter. Fig. S8.2 depicts how the 2D embedding (by t-SNE) of the $\nu = 22$ dimensional latent representations is correlated with the molecule diameter and cavity diameter of a porous organic cage molecule.

Figure S8.1: The first component of the latent representation of a cage is strongly correlated to its cavity diameter. The cavity diameter was computed by pywindow. In the reconstruction of the 3D porosity image of a given cage, the first component of the latent representation (in the first column of $U_\nu \Sigma_\nu$) is the weight on the first eigencage $v_1$ in the reconstruction of the 3D porosity image via eqn. 8. That the first eigencage is a good descriptor of pore size is suggested by the radial symmetry of the core of the first eigencage in Fig. 5 that is added to the approximately radially symmetric average cage $\bar{c}$ to reconstruct the 3D porosity image via eqn. 8.
Figure S8.2: The 2D embedding (by t-SNE) of the learned latent representation of the 3D porosity images contained in the rows of $U_\nu \Sigma_\nu$. Here, the diameter of points represents molecule diameter; color represents cavity diameter, both computed by pywindow.\(^1\) Note that neighbors in latent space tend to have similar cavity and molecule diameters.
S9  Relationship of latent representation and number of windows to enter cavity

Figure S9.1: The 2D embedding (by t-SNE) of the learned latent representation of the 3D porosity images contained in the rows of $U_{\nu} \Sigma_{\nu}$. Here, the color represents the number of windows possessed by the cage molecule. The gas molecules enter the cavity of the cage through the windows. The number of windows is computed by `pywindow`. Note that cages within clusters in latent space have the same number of windows.
S10 Assessing effects of flexibility in cages

To evaluate how sensitive our methods are to flexibility, four cages (CC2, CC3, CC4 and CC5) were allowed to fluctuate at 298 K in a molecular dynamics (MD) simulation. We considered an isolated cage only. The simulations were performed using GULP v5.0\textsuperscript{10} in the NVT ensemble using the leapfrog Verlet algorithm with the Nose-Hoover thermostat following an initial geometrical optimization of the cage structure. The cage-specific force field (CSFF) was used to model the bond, angle and torsion potentials;\textsuperscript{11} missing parameters were either obtained from the polymer consistent force field (PCFF)\textsuperscript{12–14} as implemented by DL\textsubscript{FIELD} v4.3\textsuperscript{15} or a slightly modified version of a PCFF parameter was used. The MD files with every force field parameter can be found on github.com/SimonEnsemble/latent_cage_space/MD_files/*.*gin, where the origin of each potential parameter can be seen in the comments. The CSFF was specifically developed to properly describe porous organic cages. Both Coulombic and Lennard Jones (12-6) potentials were used for interactions between atoms separated by two or more atoms in the bonding graph. Partial charges were assigned to the cage atoms using the parameters provided by the PCFF as given by Holden et al.\textsuperscript{11} No periodic boundaries were used in these simulations since we considered isolated cage molecules. The equilibration time was 50 ps while the production time was 1 ns with a timestep of 0.5 fs.

We initiated the MD simulation on the aligned cages (as opposed to cage structures taken directly from Miklitz et al.\textsuperscript{16}) rather than going through the process of re-aligning the MD snapshots, due to the excessive computational time required for the Coherent Point Drift algorithm. During the MD simulation, 400 snapshots of each fluctuating cage were gathered. To help confirm the validity of the simulations, we calculated the window diameter for each snapshot in CC3 using pywindow (Fig. S10.1). Holden et al.\textsuperscript{11} did a similar analysis of the CC3 window diameter during a MD simulation with the before-mentioned CSFF force field, and our result appears to be in good agreement with theirs.

We computationally scanned each MD snapshot of the fluctuating cages to generate a 3D
Figure S10.1: A histogram of the window diameters, calculated using \texttt{pywindow}, exhibited by each snapshot of \textbf{CC3} from molecular dynamics simulations described above. All four windows of \textbf{CC3} are included.

porosity image \( \mathbf{c}_i \) for each snapshot \( i \). We then subtracted off the mean 3D porosity image among the 74 rigid-cage dataset, \( \bar{\mathbf{c}} \) to yield \( \mathbf{a}_i \). To visualize where the 3D porosity images of the fluctuating cages are located in latent space, we do not use t-SNE. If t-SNE were used, we would have to retrain it with the inclusion of the fluctuating cages, which would result in overfitting to the large number of fluctuating cages. That is, t-SNE does not learn an explicit mapping function to project new, unseen data points— the 3D porosity images of the MD snapshots— onto the embedding in Fig. 7. To illustrate how flexible cages move around latent space as their cavity shapes fluctuate, we adopt the simpler choice to project the cages onto our learned latent space \( \mathbf{U}_\nu \Sigma_\nu \) on the basis of the 74 rigid-cage molecules in Fig. 1, but then visualize only the first two latent dimensions. This is equivalent to using a \( \nu = 2 \) dimensional approximant in eqn. 6, then projecting the MD snapshots of flexible cages onto the \( \nu = 2 \) dimensional latent space. The idea here is to express each 3D porosity image of a snapshot, \( \mathbf{a}_i \), as a linear combination of the first two eigencages:

\[
\mathbf{a}_i^T = \ell_1 \mathbf{v}_1^T + \ell_2 \mathbf{v}_2^T.
\] (S3)
The weights \( \ell := [\ell_1, \ell_2] \) form the projection of \( \mathbf{a}_i \) onto the \( \nu = 2 \)-dimensional latent space. We can solve for the weights as \( \mathbf{a}_i^T \mathbf{V}_2 \), where \( \mathbf{V}_2 = [\mathbf{v}_1, \mathbf{v}_2] \) contains the first two eigencages in its columns.

Fig. S10.2 shows the projections of the MD snapshots in addition to the first two dimensions of the latent representations \( \mathbf{U}_2 \Sigma_2 \) of the original 74-rigid-cage dataset. All four cages undergo thermal fluctuations and explore a small region of the latent space. In the rigid view of cages, we think about each cage as a point in latent space. In the flexible view of cages, each cage explores a region in latent space. The size and shape of the region is dictated by the distribution of cavity shapes adopted while undergoing thermal fluctuations. Out of the four selected cages, CC5 is the largest and most flexible, leading to a larger region of latent space explored. Comparatively, MD snapshots of CC2, CC3 and CC4 occupy a much smaller region than CC5 due to their smaller size and more rigid structure. Indeed, the histogram of pore diameters exhibited by the snapshots in Fig. S10.3 confirms that CC5 explores a broader range of pore sizes than each CC2, CC3, and CC4 during the MD simulations.

Note that the region explored by the MD snapshots in Fig. S10.2 does not overlap the latent representation of the rigid structure pulled from Miklitz et al.\textsuperscript{16} (correspondingly colored solid point). The reason is that force field predicts a slightly different minimum-energy structure than the cage pulled from the Cambridge Structural Database by Miklitz et al.\textsuperscript{16} Regardless, Fig. S10.2 provides valuable insight, showing how flexible cages explore regions of latent space as their cavity fluctuates.
Figure S10.2: The 3D porosity images of the snapshots of the flexible cages CC2 (cyan), CC3 (orange), CC4 (purple) and CC5 (maroon) during molecular dynamics simulations are projected onto the first two dimensions of our latent space. The green points are the first two dimensions of the latent representations $U_2 \Sigma_2$ of the 74 rigid cages in Fig. 1. Each fluctuating cage explores a latent region. The region explored is determined by the different geometrical configurations its cavity explores while undergoing thermal fluctuations. (The region explored by the MD snapshots does not overlap the latent representation of the rigid structure pulled from Miklitz et al.\textsuperscript{16} (solid point) because the molecular model (CSFF/PCFF) predicts a slightly different minimum-energy structure than the cage pulled from the Cambridge Structural Database by Miklitz et al.\textsuperscript{16}) Note MC6 is omitted since it is an outlier and stretched the axes on the plot.
Figure S10.3: A histogram of the pore diameter, calculated using \texttt{pywindow}, exhibited by snapshots during the MD simulations. The colors denote the histograms for CC2, CC3, CC4, and CC5.
S11 Exploring accessibility of cages to xenon

During the Henry coefficient calculations, we insert xenon atoms at random positions within a simulation box encompassing the cage molecule (see Sec. S7). These simulations may insert a xenon atom internal to a cage cavity whose window is too small for xenon to enter. This is an undesirable scenario, since xenon could practically never diffuse into the cavity in practice, yet that inner cavity is contributing to the simulated Henry coefficient. Here, we determine which cages have cavities that are inaccessible to a xenon or krypton atom.

To assess if an adsorbate can enter through a window into the cavity of the porous organic cages, we superimposed a grid of points over each cage molecule such that the distance between each adjacent datapoint was 0.1 Å. The potential energy of the adsorbate (xenon, krypton) was calculated at each point using a 12-6 Lennard Jones potential. Force field parameters were taken from the Universal Force Field (UFF) and geometric mixing rules were applied (14 Å cutoff). After the potential energy grid was computed, we assigned an integer value to each point, $-1$ if the energy was above a certain energy value (here $15k_BT$ where $k_B$ is the Boltzmann constant and the temperature was taken to be $T = 298$ K), which corresponds to a non-occupiable point, and $1$ if the energy was below that value, corresponding to an occupiable point. The potential energy value $15k_BT$ is defined to be a potential energy barrier that is prohibitively high for xenon to occupy. The next we see what occupiable points are connected. For every set of contiguously neighboring occupiable points, we relabel them unique values corresponding to a segment of occupiable points by applying a flood fill algorithm. If the center of the cage is occupiable and belongs to a different segment than the region outside of the cage, we deem that cage’s cavity as inaccessible. If this were the case, the xenon adsorbate would have to overcome a prohibitively large energy barrier to percolate through the window and enter the cavity. This catches the undesirable case where Henry coefficients from the simulations have unrealistic contributions from the internal cage cavity despite the adsorbate being unable to percolate into the cavity because the window diameters are too small. In Fig. 8, we mark this subset of cages with xenon-
inaccessible cavities with an ‘x’.

This method was applied to all of the 74 rigid porous organic cages, not taking into consideration that the window can “breathe”, i.e. fluctuate to allow the adsorbate to enter.\textsuperscript{16} The set of cages deemed to harbor inaccessible cavities to xenon is in Table S3, where the inaccessible cages can be seen along with their maximum window diameter (calculated with pywindow). Note that these xenon-inaccessible cages have maximum window diameters less than the diameter of a xenon atom, which is c.a. 4 Å, consistent with our method.

Miklitz et al.\textsuperscript{16} conducted a similar analysis where they used molecular dynamics (MD) to observe the pore breathing and calculated the distribution of pore-limiting envelopes (PLE) of the MD snapshots. Their dynamic PLE analysis of what percentage of the time cages have windows large enough for Xe to percolate into the cavities contrasts with our method of focusing on a static cage structure and making a binary classification as accessible or inaccessible based on the potential energy barrier at the window. Of the subset of 12 cages they studied, CP1 and WC4 were the only two they identified as having cavities inaccessible to xenon practically for all of the time (0.2% of the time); our analysis identifies CP1 and WC4 as inaccessible as well. A notable exception was RCC3a, which was open only 1.1% of the time in their MD simulations, but, according to our analysis, is accessible. A possible source of discrepancy is that the force field used in the MD simulations of Miklitz et al.\textsuperscript{16} may provide a different “average” structure than deposited in the Cambridge Structural Database (the source of our RCC3a structure from Ref.\textsuperscript{16}). Other cages such as CC3, CC2 and NC2 were open to Xe roughly 10% of the time, considered accessible as in our analysis.
Table S3: A list of the cages found to be inaccessible to xenon alongside their maximum window diameter (calculated with pywindow).

| Cage   | Maximum window diameter [Å] |
|--------|-----------------------------|
| CB5    | 2.334                       |
| CP1    | 2.392                       |
| CP3    | 2.679                       |
| CP4    | 1.495                       |
| CP5    | 1.474                       |
| HC1    | 3.594                       |
| RCC1a  | 2.016                       |
| RCC1b  | 2.180                       |
| RCC1c  | 2.786                       |
| RCC1d  | 2.578                       |
| WC2    | 2.682                       |
| WC3    | 1.578                       |
| WC4    | 3.281                       |
S12 Adsorption site analysis

Here, we analyze the distribution of adsorption sites of xenon internal and external to the cavity of the cage molecules. We consider any local minimum in the potential energy as a potential adsorption site. We identified potential energy minimia on the basis of a computed potential energy grid as in Sec. S11. We looked at the 3D array storing the potential energy grid, looped over each element, and, if no lower values were adjacent to it, stored it as a local minimum. All local minima exhibiting positive energy values (repulsive) were discarded. Fig. S12.1 shows each local energy minimum as a point; the $y$-axis is the value of the potential energy at the local minimum and the $x$-axis is its distance from the center of mass of the cage. For reference, we plotted as a vertical line the radius of the cavity, computed by taking the 2-norm of the vector between the center of mass of the cage and the closest atom and correcting it with the van der Waals radius of the closest atom. This allows us to assess the relative potential energies of the adsorption sites internal to the cavity versus on the periphery or window site of the molecule.
Figure S12.1: Energy of each local potential energy of xenon minima [kJ mol\(^{-1}\)] plotted against the distance from the center of mass of each cage [Å]. The red frames correspond to the cages deemed inaccessible by our flood fill algorithm. The cavity diameter is plotted as a vertical, dotted line to give an idea of the relative potential energies of adsorption sites within the cavity versus on the periphery or window site of the cage.
S13 Visualization of the cage structures

Larger images of the cages in Fig. 1 are shown in Fig. S13.1.
CC10
CC1
CC2
CC3
CC4
CC5
CC9
CD1
CD2
CD3
CP1
CP3

S33
Figure S13.1: Visualizations of the structures of all 74 porous organic cage molecules analyzed in this study. The .xyz files of these cages are from Miklitz et al. and Greenaway et al., see these references for the naming conventions. The cages are visualized in their centered and aligned states in the $[-20, 20]^3$ Å snapshot box in preparation for taking the 3D cage cavity image.
S14 Safety statement

As this research is all computational in nature, no unexpected or unusually high safety hazards were encountered.
References

(1) Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *Journal of Chemical Information and Modeling* 2018,

(2) Widnall, S.; Deyst, J.; Greitzer, E. Lecture L26: 3D Rigid Body Dynamics: The Inertia Tensor. *MIT Open Courseware: Dynamics 16.07* 2008,

(3) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008, 9, 2579–2605.

(4) Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill* 2016,

(5) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society* 1992, 114, 10024–10035.

(6) Frenkel, D.; Smit, B. *Understanding molecular simulation: from algorithms to applications*; Elsevier, 2001; Vol. 1.

(7) PorousMaterials.jl. 2018; [http://dx.doi.org/10.5281/zenodo.1400839](http://dx.doi.org/10.5281/zenodo.1400839).

(8) Patil, R. S.; Banerjee, D.; Simon, C. M.; Atwood, J. L.; Thallapally, P. K. Noria: A Highly Xe-Selective Nanoporous Organic Solid. *Chemistry–A European Journal* 2016, 22, 12618–12623.

(9) Chen, L. et al. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nature Materials* 2014, 13, 954.

(10) Gale, J. D. GULP: A computer program for the symmetry-adapted simulation of solids. *J. Chem. Soc., Faraday Trans.* 1997, 93, 629–637.
(11) Holden, D.; Jelfs, K. E.; Cooper, A. I.; Trewin, A.; Willock, D. J. Bespoke Force Field for Simulating the Molecular Dynamics of Porous Organic Cages. *The Journal of Physical Chemistry C* **2012**, *116*, 16639–16651.

(12) Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T. An ab initio CFF93 all-atom force field for polycarbonates. *Journal of the American Chemical Society* **1994**, *116*, 2978–2987.

(13) Sun, H. Force field for computation of conformational energies, structures, and vibrational frequencies of aromatic polyesters. *Journal of Computational Chemistry* **1994**, *15*, 752–768.

(14) Sun, H. Ab initio calculations and force field development for computer simulation of polysilanes. *Macromolecules* **1995**, *28*, 701–712.

(15) Yong, C. W. Descriptions and Implementations of DL_F Notation: A Natural Chemical Expression System of Atom Types for Molecular Simulations. *Journal of Chemical Information and Modeling* **2016**, *56*, 1405–1409, PMID: 27455451.

(16) 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. *The Journal of Physical Chemistry C* **2017**, *121*, 15211–15222.

(17) Kim, J.; Martin, R. L.; Rulíbel, O.; Haranczyk, M.; Smit, B. High-throughput characterization of porous materials using graphics processing units. *Journal of chemical theory and computation* **2012**, *8*, 1684–1693.

(18) Martin, R. L.; Prabhat.; Donofrio, D. D.; Sethian, J. A.; Haranczyk, M. Accelerating analysis of void space in porous materials on multicore and GPU platforms. *The International Journal of High Performance Computing Applications* **2012**, *26*, 347–357.
(19) Greenaway, R.; Santolini, V.; Bennison, M.; Alston, B.; Pugh, C.; Little, M.; Miklitz, M.; Eden-Rump, E.; Clowes, R.; Shakil, A.; Cuthbertson, H.; Armstrong, H.; Briggs, M.; Jelfs, K.; Cooper, A. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nature Communications* **2018**, *9*, 2849.
