Eigencages: Learning a Latent Space of Porous Cage Molecules

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

School of Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, Oregon, United States

Supporting Information

ABSTRACT: Porous organic cage molecules harbor nanosized 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 three-dimensional (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 congregate. 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 by approximately 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. Our methods could be applied to learn latent representations of cavities within other classes of porous materials and of shapes of molecules in general.

INTRODUCTION

More than 10% of the world’s energy consumption is devoted to purifying chemical mixtures.¹ 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,² instead via selective adsorption on a solid-state, porous material.³ Porous solids composed of porous organic cage molecules⁴ have auspiciously demonstrated the ability to separate gases⁵ in application to carbon dioxide capture from natural gas⁶ and flue gas of coal-fired power plants,^{7,8} xenon/krypton separations,^{9,10} and sulfur hexafluoride capture.¹¹ Porous cage solids with high adsorptive selectivities may serve as gas sensors as well.^{12,13}

Porous organic cages^{5,14,15} 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.¹⁶ Unlike metal-¹⁷ and covalent-¹⁸ 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 noncovalent/noncoordination interactions between them.¹⁹ On the order of hundreds of porous organic cages have been reported.²⁰

For deployment in molecular separations or sensing, the size and shape of the cavities in a porous material can strongly influence the adsorptive selectivity.^{21–25} 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.²⁴ 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,²⁶ and the entropy of adsorption, e.g., by minimizing the loss of rotational entropy in the cavity.²⁷ Therefore, it is important to 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,^{28,29} algebraic topology,³⁰ radial distribution functions,³¹ and density of minima in the potential energy landscape.³² For extended networks, the MOFomics³³ 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.³⁴

Received: September 10, 2018

Published: December 13, 2018

Figure 1. Structures of all aligned 74 porous organic cage molecules^{16,37} comprising the training set in this study, ordered by cage molecule diameter computed from pywindow.³⁴ Note the diversity of cavities.

In this work, our goal is to map porous organic cage molecules to a lower-dimensional latent space^b 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 26 and 35; see Le et al.³⁶ 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 data set of 74 porous organic cage molecules.^{16,37} We achieve this by first computationally scanning the cage molecules to generate three-dimensional (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 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,^{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 lie. A low-dimensional *latent representation* follows by approximately expressing each cage as a combination of the *eigencages*. Visualizing a two-dimensional (2D) embedding of our *latent cage space* with t-SNE⁴¹ 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 DATA SET

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.,¹⁶ who extracted cage molecules from crystal structures determined by X-ray diffraction studies and deposited in the Cambridge Structural Database⁴² by various research groups.^{5,43,44} The remaining cages are from a recent study that employed robots and computational predictions of cage topologies³⁷ 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.³⁷ Figure 1 illustrates the diversity of cage structures comprising the data set; Figure S13.1 shows larger images of each cage. The spread of cavity and molecule diameters among the cages is shown in Figure S1.1.

SCANNING THE CAGES TO GENERATE THREE-DIMENSIONAL 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 molecule. Figure 2 depicts example raw 3D porosity images.

Figure 2. Example 3D porosity images. (a–f) The molecular structure of a cage (name and experimental reference^{37,45–47} in subcaption) is shown along with a contour (0.5) of the 3D porosity image (orange). Bounding box shows $[-20 \text{ \AA}, 20 \text{ \AA}]^3$ dimension of the snapshot, consistent for all cages.

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 $T = 298 \text{ K}$ 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 Field⁴⁸ (geometric mixing rules, cutoff radius $14 \text{ \AA}$). We use PorousMaterials.jl v0.1.3⁴⁹ 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.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 toward the center of the cage to emphasize the inner cavity. See Sec. S2.1 for more details and Figure S2.1 for example porosity point clouds.

We use the rotational dynamics of the porosity point clouds 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 nonincreasing 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 Figure 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 principal moments of inertia of its porosity point cloud differ by more than 1%, the case only for 18/74 cages (see Table 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.⁵¹

We employed Coherent Point Drift⁵¹ 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, 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 variance of the Gaussians; 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 lies 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, aligned 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 spans 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 lie 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^{40,52,53} to learn from the set of 3D porosity images of the cages in Figure 1 an approximate, lower-dimensional subspace in which the 3D porosity images of porous cage molecules lie. The *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 Figure 1. The latent representation of a cage follows from best-approximately expressing its 3D porosity image as a linear combination of the eigencages, or, equivalently, projecting its 3D porosity image onto the subspace spanned by the eigencages.

The Data Matrix, $\mathbf{A}$. We encapsulate all $c = 74$ 3D porosity images of porous cages into a data matrix $\mathbf{A}$. We first flatten the $g \times g \times g$ ($g = 50$) images into a set of raw vector representations $\{c_1, c_2, \dots, c_c\}$, 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 Figure 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 $\mathbf{A}$ 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 ($c \ll g^3$), the data matrix $\mathbf{A}$ is very wide (see Figure S3.1).

The Singular Value Decomposition (SVD). The singular value decomposition (SVD)^{40,52,53} enjoys use in genomics,⁵⁴

Figure 3. Full singular value decomposition of $\mathbf{A}$ in eq 1. Relevant rows/columns are colored and labeled.

recommender systems,⁵⁵ and image processing.⁴⁰ We provide a pedagogical overview of the SVD for the reader's convenience. We then reason that the SVD of our data matrix $\mathbf{A}$ identifies an approximate lower-dimensional subspace in which the 3D porosity images lie (the latent space) and a latent representation of the porosity of each cage molecule.

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (1)$$

where $\mathbf{U}$ is a $c \times c$ orthogonal matrix, $\mathbf{\Sigma}$ is an $c \times g^3$ diagonal matrix with the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_c \geq 0$ of $\mathbf{A}$ arranged down the diagonal, and $\mathbf{V}$ is a $g^3 \times g^3$ orthogonal matrix. The columns of $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_c]$ form an orthonormal basis for $\mathbb{R}^c$ and are the *left singular vectors* of $\mathbf{A}$; the columns of $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{g^3}]$ form an orthonormal basis for $\mathbb{R}^{g^3}$ and are the *right singular vectors* of $\mathbf{A}$. These left and right singular vectors are eigenvectors of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$, respectively; the singular values are the square roots of the [shared] non-negative eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. Since $\mathbf{A}^T\mathbf{A}$ is proportional to the sample covariance matrix, we are essentially conducting principal component analysis,⁵³ identifying the most significant directions of variance among the 3D porosity images.

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

$$\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T \quad (2)$$

where now $\mathbf{\Sigma}_r$ is a diagonal $r \times r$ matrix that contains only the r nonzero singular values of $\mathbf{A}$ down its diagonal in a nonincreasing order, $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_r]$ is a $c \times r$ matrix, and $\mathbf{V}_r = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r]$ is a $g^3 \times r$ matrix (retaining only the first r singular vectors). Writing eq 2 using an outer product expansion expresses $\mathbf{A}$ as a sum of rank-one matrices:

$$\mathbf{A} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (3)$$

The singular values appear in eq 3 as weights on the rank-one matrices, formed by the outer product of corresponding left and right (unit) singular vectors, used to construct the matrix $\mathbf{A}$. This emphasizes that the singular vectors are ordered in terms of significance.

We compute the singular value decomposition using the *svd* function in Julia.⁵⁶

A Geometric View of SVD. A useful geometric view of the SVD follows by considering how $\mathbf{A}$ maps a unit hypersphere in $\mathbb{R}^{g^3}$ into $\mathbb{R}^c$. We can express any point $\mathbf{x}_s \in \mathbb{R}^{g^3}$ on the unit

hypersphere as a linear combination of the right singular vectors:

$$\mathbf{x}_s = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \dots + \alpha_{g^3} \mathbf{v}_{g^3} \quad (4)$$

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

$$\mathbf{A}\mathbf{x}_s = \alpha_1 \sigma_1 \mathbf{u}_1 + \alpha_2 \sigma_2 \mathbf{u}_2 + \dots + \alpha_r \sigma_r \mathbf{u}_r \quad (5)$$

This follows from $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ for $i = 1, 2, \dots, r$ and $\mathbf{A}\mathbf{v}_i = 0$ for $i = r + 1, \dots, g^3$. eq 5, together with the constraint on the α_i 's, describes an r -dimensional ellipsoid lying in $\mathbb{R}^c$, whose principal semiaxes are in the direction of $\mathbf{u}_i$ with lengths of σ_i for $i = 1, \dots, r$. Therefore, multiplication by $\mathbf{A}$ 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 lies in $\mathbb{R}^c$. The SVD recovers the directions of the principal semiaxes of this ellipsoid, $\{\mathbf{u}_i\}$, and their lengths, $\{\sigma_i\}$, as well as each vector $\mathbf{v}_i$ that is mapped to $\sigma_i \mathbf{u}_i$. The set of vectors $\{\mathbf{u}_i\}$ is orthonormal and, remarkably, so is the set of vectors $\{\mathbf{v}_i\}$.

Low-Rank- ν Approximation $\mathbf{A}_\nu$ to the Data Matrix $\mathbf{A}$. Now, we employ the SVD to compress the data matrix $\mathbf{A}$ by finding a low-rank approximation. We define the “best” rank $\nu < r$ approximation to $\mathbf{A}$, $\mathbf{A}_\nu$, as the one where $\|\mathbf{A} - \mathbf{A}_\nu\|_F$ is minimized, where $\|\cdot\|_F$ is the Frobenius norm. One can show that the optimal rank- ν approximator is

$$\mathbf{A}_\nu = \sum_{i=1}^{\nu} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (6)$$

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

$$\frac{\|\mathbf{A} - \mathbf{A}_\nu\|_F}{\|\mathbf{A}\|_F} = \sqrt{\frac{\sum_{i=\nu+1}^r \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}} \quad (7)$$

From a geometric standpoint, the relative error is related to the lengths of the principal semiaxes that we collapse in approximating the r -dimensional ellipsoid in eq 5 with a ν -dimensional ellipsoid and how they compare to the longest principal semiaxes retained.

Interpreting the SVD for 3D Porosity Images of Porous Cage Molecules. The data matrix $\mathbf{A}$ encapsulates the 3D

Figure 4. Low rank approximant $A_\nu \approx A$ in eq 6. Compare to Figure 3. Eigencage k is row k of V_ν^T . The latent representation of cage i is row i of $U_\nu \Sigma_\nu$.

porosity images of all 74 porous cage molecules. We showed that we can approximate the data matrix A with a lower-rank approximant A_ν in eq 6. The right singular vectors $\{v_i\}$ lie in the space of all 3D porosity images. As only the first ν right singular vectors appear in the approximant in eq 6, the best ν -dimensional subspace of 3D porosity images is thus spanned by the orthonormal set of vectors $v_1, v_2, \dots, v_\nu$. Analogous to eigenfaces,^{39,40} we declare this set of vectors, which are fictitious 3D porosity images [well, with $\bar{c}$ subtracted and normalized to have magnitude unity] learned by SVD, the *eigencages*. Algebraically, eq 6 approximates the flattened 3D porosity image of cage k as

$$c_k^T \approx \bar{c}^T + \sum_{i=1}^{\nu} \sigma_i u_i[k] v_i^T \quad (8)$$

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

We choose the dimension of the latent space ν as the smallest such that the relative error in eq 7 is less than 15%, leading to $\nu = 22$. See Figure 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 Figure S4.1.

To summarize, the singular value decomposition, in an unsupervised manner, learns:

- the best approximate ν -dimensional subspace of $\mathbb{R}^{g^3}$ in which the 3D images of the porosity of cage molecules lie. This subspace is spanned by the set of orthonormal right singular vectors $\{v_1, v_2, \dots, v_\nu\}$, ranked in terms of importance, which we declare as *eigencages*.
- a ν -dimensional *latent space* of porous organic cage molecules defined by the weights used to best-

approximately construct a 3D porosity image from a linear combination of the eigencages. The ν -dimensional *latent representation* of cage i is the i th row of $U_\nu \Sigma_\nu$.

The latent representation of a new cage absent from Figure 1 can be obtained by, instead of building a new matrix A with an additional row and recomputing the SVD, projecting its 3D porosity image onto the subspace spanned by the eigencages $\{v_1, \dots, v_\nu\}$.

The Eigencages. The eigencages, the rows of V_ν^T (see Figure 4), are an orthonormal basis for the approximate lower dimensional subspace in which all 3D porosity images lie 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 data set.⁵³ We visualize the first (and most important) six eigencages in Figure 5. As eq 8 illustrates, the eigencages express deviations from the average 3D porosity image $\bar{c}$, whose contours are shown in Figure 5a. Thus, the eigencages possess contours at both positive and negative values. The periphery of the first eigencage v_1 in Figure 5b exhibits radial symmetry; as Figure S8.1 indicates, v_1 is used heavily to describe how the cavity diameter of a given cage differs from the average cage in Figure 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 eq 8). We show here that, remarkably, even the six eigencages in Figure 5 can be enough to visually reproduce the structure of the cavity in a porous organic cage molecule. As an example, in Figure 6 we show the approximate reconstruction of the 3D porosity image of cage molecule B25 using different numbers of eigencages. Mathematically, Figure 6 shows the approximant given by eq 8 with $\nu = 1, 2, \dots, 9$. Expressing B25 in terms of the first two eigencages is insufficient to capture its shape; for $\nu = 4$, the cavity and windows of B25 are captured to some extent. Only until we reach $\nu = 6$ does the outer-contour of the reconstructed 3D void space image of B25 well-approximate

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).

the shape of the molecule. The reconstructed **B25** 3D porosity image with $\nu = 9$ is approaching visual indistinguishability from the exact 3D porosity image in Figure 6a. This shows that, e.g., a nine-dimensional latent representation ($\nu = 9$) of **B25** contains sufficient information to visually reproduce the shape of its cavity.

VISUALIZING AND 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 $U_\nu \Sigma_\nu$ (see Figure 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)^{41,57} to embed our learned latent representations of porous cages into a 2D space for visualization. In contrast to simply visualizing the first two

Figure 6. Reconstructing cage **B25** with its latent representation. (a) Exact 3D void space image of **B25**. (b–j) Reconstructions using latent representations of varying dimensions ν . These are contours (0.5) of eq 8 with varying ν (views from an angle to clearly see the cavity).

dimensions of our latent space (choosing $\nu = 2$ in eq 6), t-SNE is a nonlinear 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.⁴¹ A tunable parameter in t-SNE is the perplexity, a loose measure of how many neighbors each data point has.^{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 41). Figure 7 shows the resulting 2D embedding of the latent representations 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 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. Figure 7 highlights a few salient clusters. The remaining clusters are shown in Figure 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 thousands of porous cages instead of manually grouping them together.

Figure 7. Latent representations of cages $U_v \Sigma_v$ embedded into 2D by t-SNE.^{41,57} Salient clusters are highlighted. See Figure S6.1 for the remaining clusters.

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 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.^c Compared to human-engineered/invented features of cages, we expect that our discovered latent space defined in $U_v \Sigma_v$ 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 Figure 6 shows, the

latent representation of a cage can visually reproduce the shape of its cavity; (iii) as Figure 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), and then subtract the Henry coefficient of helium to mimic an excess adsorption experiment, as in Patil et al.¹⁰ See Sec. S7 for details, discussion, and comparison to experimental Xe and Kr adsorption measurements in noria (NC2)¹⁰ and CC3.⁹ The

Figure 8. Latent representations of cages $U_i \Sigma_i$ embedded into 2D by t-SNE.^{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.

Figure 9. Walking through latent space from cage NC2 (a) to DC1 (f). Panels (b–e) are fictitious cage cavities generated by walking along a line between the latent representation of NC2 and DC1.

points in Figure 8 are colored according to the simulated Xe/Kr selectivity at infinite dilution. A subset of the cages, marked as 'X's in Figure 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 diameters of the cages computed by pywindow³⁴ are correlated with the location of a cage in latent space. Figure 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, Figure 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, Figures 8, S8.2, and S9.1 show that our latent representations of porous cage molecules are useful for predicting properties that are correlated with the shape of the cavity (as is the case for Xe/Kr selectivity^{25,26}) and thus lend a meaningful notion of similarity between two cage molecules.

Figure 10. Flexible cages explore an area in latent space. (a) Snapshots of the structure of an isolated, empty cage CC5 undergoing thermal fluctuations during a molecular dynamics simulation. (b) The gray points are the $\nu = 2D$ latent representations of the (rigid) cages learned by SVD, rows of $U_2 \Sigma_2$. The 3D porosity images of 400 snapshots of cages CC2, CC3, CC4, and CC5 (colored) during molecular dynamics simulations are projected onto this 2D latent space. To precisely define the axes, the 3D porosity image of cage k in the scatter plot is best-approximated in terms of a linear combination of the first two eigencages as $\mathbf{c}_k \approx \bar{\mathbf{c}} + l_{1,k} \mathbf{v}_1 + l_{2,k} \mathbf{v}_2$. Each fluctuating cage explores a latent region determined by the set of cavity shapes explored while undergoing thermal fluctuations. MC6, an outlier, is omitted to avoid stretched axes.

A Walk through Latent Cage Space. Gómez-Bombarelli et al.⁵⁹ 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”.⁶⁰ Similarly, we show in Figure 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,^{9,61–63} and the shape of their cavity fluctuates,⁹ which influences adsorption.⁶⁴ 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 the polymer consistent force field (PCFF)^{65–67} as implemented by DL_FIELD v4.3⁶⁸ and the cage specific force field (CSFF).⁶⁹ We gathered snapshots of each cage while they were fluctuating in the MD simulation, computed their 3D porosity images, and then projected them onto a $\nu = 2D$ latent space. See Sec. S10 for details. Figure 10 illustrates that flexible cages, on the basis of the MD snapshots, each occupy a region in latent space—instead of a point, as in the static view of cages. 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.⁷⁰ Since then, advanced classes of materials harboring nanosized pores such as porous cage solids^{5,14} have emerged and offer high adsorptive selectivities.^{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,²¹ it is important to mathematically characterize the geometry of 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 3D 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 congregate in a much lower dimensional subspace of this enormous space. Using the singular value decomposition, we learned in an unsupervised manner the approximate lower-dimensional subspace in which the void spaces of porous organic cages sit, which is characterized by a set of *eigencages*. Expressing a cage approximately 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 2D 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 tend 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 Figure 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 Figure 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,⁷¹ our latent space will be useful for 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^{9,61–63} as opposed to rigid as we took the cages here. In reality, we showed in Figure 10 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.⁵ In fact, assembly of porous cage molecules to form a bulk solid can be sensitive to small changes in the cage molecule.⁷² 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!⁴ For this reason, the simulated Xe/Kr selectivities in the isolated cage molecules in Figure 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 Figure S12.1, however, suggests that the dominant adsorption site in cages exhibiting the highest simulated Xe/Kr selectivity tends to be internal to the cavity; hence false positives may be unlikely unless the internal cavity sites are blocked in the bulk solid. 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,⁴ 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 Figure 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 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 intercage 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 rotation-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-SNE⁴¹ 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.⁴¹ Our reasons for employing SVD instead of t-SNE directly are (i) the mapping by SVD lends more interpretability through visualizing the eigencages (Figure 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 in Figure 10 and (b) generate fictitious pore structures as we did between NC2 and DC1 in Figure 9, (iii) t-SNE is stochastic, whereas SVD is deterministic.

To further the ideas in this study, we are working toward 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 reported⁷³ in comparison to the hundreds of porous cage solids;²⁰ 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, nonlinear dimensionality reduction algorithms such as autoencoders⁷⁴ with convolutional layers^{75,76} may fare at learning/detecting hierarchical features of pore structures with translational and rotational invariance.

■ ASSOCIATED CONTENT

■ Supporting Information

A Supporting Information document is also available. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.8b00638.

A Jupyter Notebook with Julia⁵⁶ 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 v0.1.3.⁴⁹ (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Cory.Simon@oregonstate.edu.

ORCID

Cory M. Simon: 0000-0002-8181-9178

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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.

■ ADDITIONAL NOTES

^aThe energy requirement for separating a mixture has a thermodynamic limit,³ of course.

^b“Latent” refers to “hidden”: abstractly, the latent space is a lower-dimensional subspace/manifold—which is embedded in data space—in/on which the data congregates.

^cAkin 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”.⁵⁸

^dWe did encourage 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.

■ 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 U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Advanced Manufacturing Office 2015.
- (3) Bhowan, A. S.; Freeman, B. C. Analysis and status of post-combustion carbon dioxide capture technologies. *Environ. Sci. Technol.* **2011**, 45, 8624–8632.
- (4) Tozawa, T.; et al. Porous organic cages. *Nat. Mater.* **2009**, 8, 973.
- (5) Hasell, T.; Cooper, A. I. Porous organic cages: soluble, modular and molecular pores. *Nat. Rev. Mater.* **2016**, 1, 16053.
- (6) Mastalerz, M.; Schneider, M. W.; Oppel, I. M.; Presly, O. A salicylbisimine cage compound with high surface area and selective CO₂/CH₄ adsorption. *Angew. Chem., Int. Ed.* **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. *Angew. Chem., Int. Ed.* **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. *Angew. Chem., Int. Ed.* **2015**, 54, 13241–13244.
- (9) Chen, L.; et al. Separation of rare gases and chiral molecules by selective binding in porous organic cages. *Nat. Mater.* **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. *Chem. - Eur. J.* **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. *J. Am. Chem. Soc.* **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. *Adv. Mater.* **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. *Chem. Commun.* **2013**, 49, 8398–8400.
- (14) Holst, J. R.; Trewin, A.; Cooper, A. I. Porous organic molecules. *Nat. Chem.* **2010**, 2, 915.
- (15) Cooper, A. I. Porous molecular solids and liquids. *ACS Cent. Sci.* **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. *J. Phys. Chem. 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, No. eaal1585.
- (19) McKeown, N. B. Nanoporous molecular crystals. *J. Mater. Chem.* **2010**, 20, 10588–10597.
- (20) Evans, J. D.; Huang, D. M.; Haranczyk, M.; Thornton, A. W.; Sumbly, C. J.; Doonan, C. J. Computational identification of organic porous molecular crystals. *CrystEngComm* **2016**, 18, 4133–4141.
- (21) Mitra, T.; Jelfs, K. E.; Schmidtman, M.; Ahmed, A.; Chong, S. Y.; Adams, D. J.; Cooper, A. I. Molecular shape sorting using molecular organic cages. *Nat. Chem.* **2013**, 5, 276–281.
- (22) Zhu, W.; Kapteijn, F.; Moulijn, J. Shape selectivity in the adsorption of propane/propene on the all-silica DD3R. *Chem. Commun.* **1999**, 2453–2454.
- (23) Lee, Y.; Barthel, S. D.; Dlotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. High-throughput screening approach for nanoporous materials genome using topological data analysis: application to zeolites. *J. Chem. Theory Comput.* **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? *Chem. Mater.* **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. *Angew. Chem.* **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. *J. Mol. Graphics Modell.* **2013**, *44*, 208–219.
- (29) Martin, R. L.; Smit, B.; Haranczyk, M. Addressing challenges of identifying geometrically diverse sets of crystalline porous materials. *J. Chem. Inf. Model.* **2012**, *52*, 308–318.
- (30) Lee, Y.; Barthel, S. D.; Dlotko, P.; Moosavi, S. M.; Hess, K.; Smit, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nat. Commun.* **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. *J. Phys. Chem. C* **2013**, *117*, 14095–14105.
- (32) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **2009**, *130*, 104504.
- (33) First, E. L.; Floudas, C. A. MOFomics: Computational pore characterization of metal–organic frameworks. *Microporous Mesoporous Mater.* **2013**, *165*, 32–39.
- (34) Miklitz, M.; Jelfs, K. E. pywindow: Automated Structural Analysis of Molecular Pores. *J. Chem. Inf. Model.* **2018**. DOI: 10.1021/acs.jcim.8b00490
- (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. *Mol. Syst. Des. Eng.* **2019**. DOI: 10.1039/C8ME00050F
- (36) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.* **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. *Nat. Commun.* **2018**, *9*, 2849.
- (38) Sirovich, L.; Kirby, M. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. 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, 1991; pp 586–591.
- (40) Muller, N.; Magaña, L.; Herbst, B. M. Singular value decomposition, eigenfaces, and 3D reconstructions. *SIAM Rev.* **2004**, *46*, 518–545.
- (41) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- (42) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.
- (43) Evans, J. D.; Sumby, C. J.; Doonan, C. J. Synthesis and applications of porous organic cages. *Chem. Lett.* **2015**, *44*, 582–588.
- (44) Zhang, G.; Mastalerz, M. Organic cage compounds—from shape-persistency to function. *Chem. Soc. Rev.* **2014**, *43*, 1934–1947.
- (45) Crini, G. A history of cyclodextrins. *Chem. Rev.* **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. *Eur. J. Org. Chem.* **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. *Chem. Commun.* **2015**, *51*, 1976–1979.
- (48) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard, W. A., III; Skiff, W. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- (49) PorousMaterials.jl, 2018; DOI: 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. *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. *Proc. Natl. Acad. Sci. U. S. A.* **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 Rev.* **2017**, *59*, 65–98.
- (57) Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill* **2016**. DOI: 10.23915/distill.00002
- (58) Box, G. E. Science and Statistics. *J. Am. Stat. Assoc.* **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 Cent. Sci.* **2018**, *4*, 268–276.
- (60) van Deursen, R.; Raymond, J.-L. Chemical space travel. *ChemMedChem* **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. *J. Phys. Chem. 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. *J. Phys. Chem. 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. *J. Am. Chem. Soc.* **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. *J. Am. Chem. Soc.* **1994**, *116*, 2978–2987.
- (66) Sun, H. Force field for computation of conformational energies, structures, and vibrational frequencies of aromatic polyesters. *J. Comput. Chem.* **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. *J. Chem. Inf. Model.* **2016**, *56*, 1405–1409.

- (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. *J. Phys. Chem. C* **2012**, *116*, 16639–16651.
- (70) Boyle, R. The Absorption of the Radioactive Emanations by Charcoal. *J. Phys. Chem.* **1907**, *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.; Schmidtman, 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. *J. Am. Chem. Soc.* **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. *Chem. Mater.* **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. *Adv. Neural Inf. Process. Syst.* **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, pp 2528–2535.