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. 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 adsorptive selectivity. For comparing cages and predicting their adsorption properties, we embed/encode the cavities of a set of 74 porous organic cage molecules into a low-dimensional, latent "cage space". We first scan the cavity of 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 cage cavity images lay. The "eigencages" are the set of orthogonal characteristic 3D cage cavity 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 hexaflouride 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 cavity 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

[†]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. 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;[?] 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 scanning the cavities of the cages to generate 3D images of their porosity. 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 cage cavity images, which belong to

an enormous space, approximately lay in a much lower-dimensional subspace. Inspired by eigenfaces in facial recognition,[? ? ?] we then employ the singular value decomposition to, in an unsupervised manner, identify a set of characteristic 3D cavity images– *eigencages*– that form an orthonormal basis for this approximate, lower-dimensional subspace in which the 3D cavity images lay. A low-dimensional *latent representation* follows by expressing each cage as a combination of the eigencages. By embedding the low-dimensional latent representations into a 2D *latent cage space* for visualization, we show that the learned encoding captures salient features of the cavities of porous cages and is predictive of simulated xenon/krypton selectivity.

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.[?] The remaining cages are from a recent study that employed robots and crystal structure prediction[?] to synthesize 33 different cages. Fig. **??** illustrates the diversity of cage structures comprising the data set; Fig. **??** shows larger images of each cage. The spread of cavity and molecule diameters among the cages is shown in Fig. **??**.

Scanning the cages to generate 3D cage cavity images

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. Fig. **??** depicts example raw 3D cage cavity images.

We first center and align the cages in a consistent manner; a meaningful interpretation of the singular value decomposition is predicated on each pixel of the 3D cage cavity image corresponding to the same relative spatial location for each cage. Centering and alignment is also required for images of human faces for eigenfaces, so that e.g. the nose and eyes appear

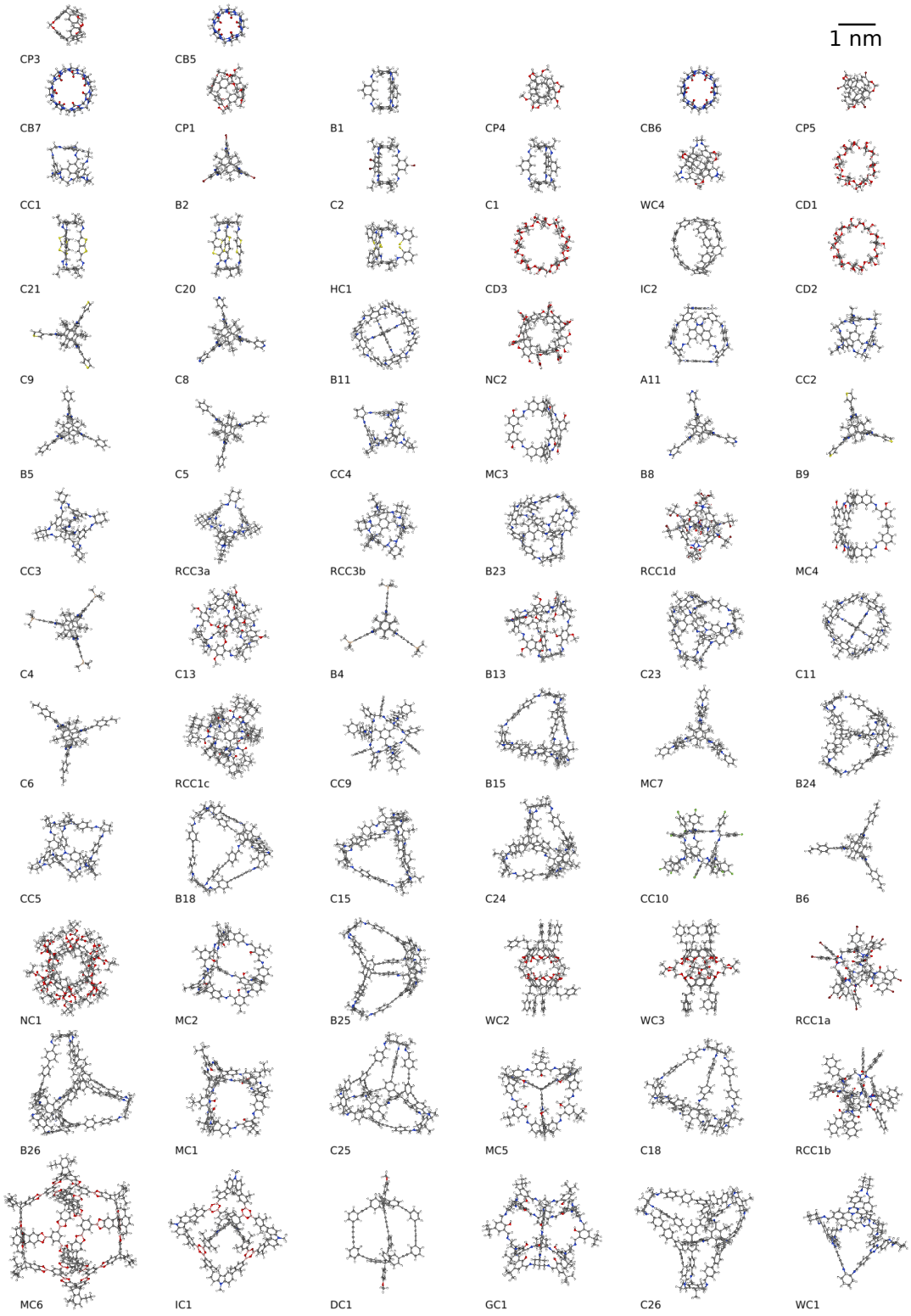
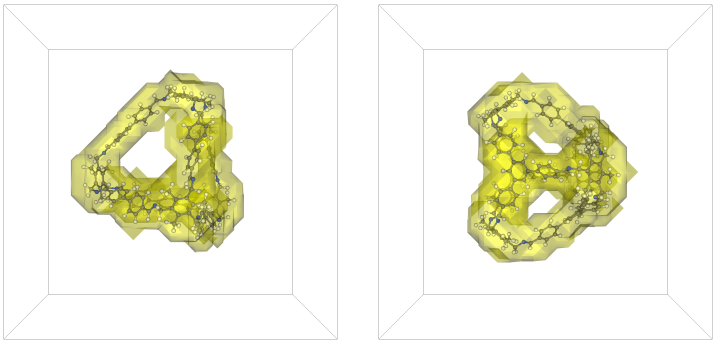
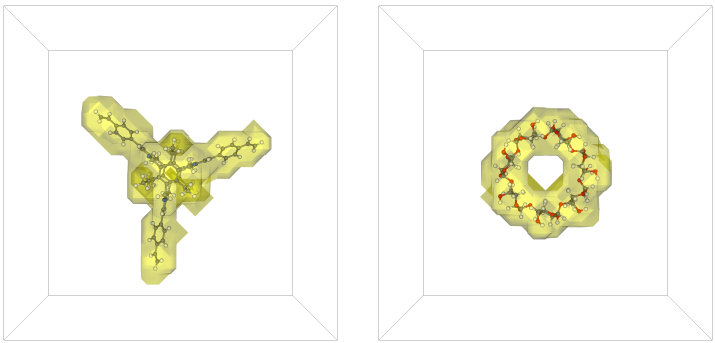


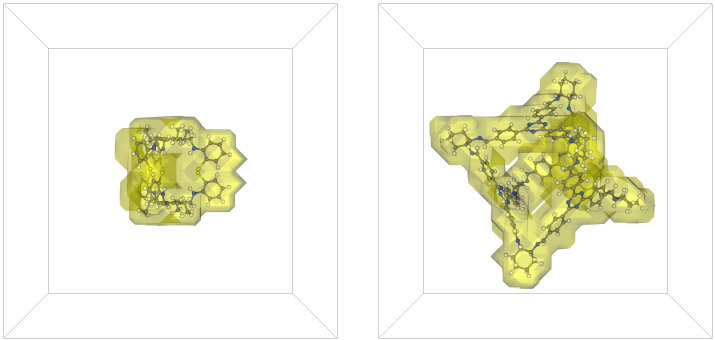
Figure 1: The structures of all 74 porous organic cage molecules[? ?] comprising the training set in this study, ordered by cage molecule diameter computed from `pywindow`.[?] Note the diversity of cavities.



(a) B15 (b) B24



(c) C6 (d) CD2



(e) HC1 (f) WC1

Figure 2: Example 3D cage cavity images. The molecular structure of the cage is shown along with a contour (0.5) of the 3D cavity image (yellow). Bounding box shows $[-20\ \text{Å}, 20\ \text{Å}]^3$ dimension of the snapshot, consistent for all cages.

in the same region for each face.[?] First, we translate each cage molecule so that its center of mass is at the origin. Second, we rotate each cage so that its principal axes of rotation are aligned with the x, y, and z Cartesian coordinate axes, with principal moments of inertia arranged in a non-increasing order. See Sec. **??** for details.

For each aligned cage, we overlay a regular, $g \times g \times g$ grid of points ($g = 25$) 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 cage cavity image* is then a 3D array; element (i, j, k) is 0 if grid point (i, j, k) is classified as void and 1 otherwise. We determine whether a given point in/around a cage is void, opposed to overlapping an atom of the cage, by computing the potential energy of a helium adsorbate at that point and assessing if the 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. 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[?] (geometric mixing rules, cutoff radius 14 Å). The potential energies are computed using `PorousMaterials.jl` v0.1.1.[?]

Learning an approximate subspace of cage cavity images

The *raw* representations of porous organic cage molecules as $g \times g \times g$ 3D cage cavity images lie in a very high-dimensional space ($g^3 = 15,625$; 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 cage cavity image by choosing each pixel as 0 or 1 in a Bernoulli process; it is extremely likely that this image will not resemble the cavity of any known cage. That is,

the effective dimension of 3D cage cavity images is much lower than $g^3 = 15,625$.

We now leverage the singular value decomposition[? ? ?] to learn from the set of 3D cage cavity images of the porous cages in Fig. **??** an approximate, lower-dimensional subspace in which the 3D cage cavity images of porous cage molecules lay. The *eigencages* are a set of orthonormal vectors that span this lower-dimensional subspace; they are ordered in terms of which directions in 3D cage cavity image space account for the most variance among the 3D cage cavity images in the data set. Expressing a 3D cage cavity 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 cage cavity images of porous cages into a data matrix $\mathbf{A}$. We first flatten the $g \times g \times g$ ($g = 25$) 3D images into a set of *raw* vector representations $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_c\}$, all of which lie in the enormous space $\mathbb{R}^{g^3}$. We compute the average 3D cage cavity image as $\bar{\mathbf{c}} = \frac{1}{c}\sum_{i=1}^{c} \mathbf{c}_i$ (visualized later in Fig. **??**). Now let $\mathbf{a}_i := \mathbf{c}_i - \bar{\mathbf{c}}$ be the difference between the void space image of cage i and the average void space image. The $c \times g^3$ data matrix $\mathbf{A}$ is then defined by assigning its ith row to be $\mathbf{a}_i^T$. As there are many fewer cages than pixels in the 3D cage cavity images ($c << g^3$), the data matrix $\mathbf{A}$ is very wide (see Fig. **??**).

The singular value decomposition (SVD)

The singular value decomposition (SVD)[? ? ?] enjoys use in genomics,[?] recommender systems,[?] and image processing.[?] We now reason that the SVD of our data matrix $\mathbf{A}$ identifies an approximate lower dimensional subspace in which the 3D cage cavity 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 $\mathbf{A} \in \mathbb{R}^{c \times g^3}$ $(g^3 > c)$ is:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{1}$$

where $\mathbf{U}$ is a $c \times c$ orthogonal matrix, $\mathbf{\Sigma}$ is an $c \times g^3$ diagonal matrix with the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_c \geq 0$ of $\mathbf{A}$ arranged down the diagonal, and $\mathbf{V}$ is a $g^3 \times g^3$ orthogonal matrix. The columns of $\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_c]$ form an orthonormal basis for $\mathbf{R}^c$ and are the *left singular vectors* of $\mathbf{A}$; the columns of $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_{g^3}]$ 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 non-zero singular values are the square roots of the [shared] non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$. Given that the latter matrix is proportional to the sample covariance matrix, we are essentially conducting principal component analysis,[?] seeking to find the most significant directions of variance among the 3D cage cavity images.

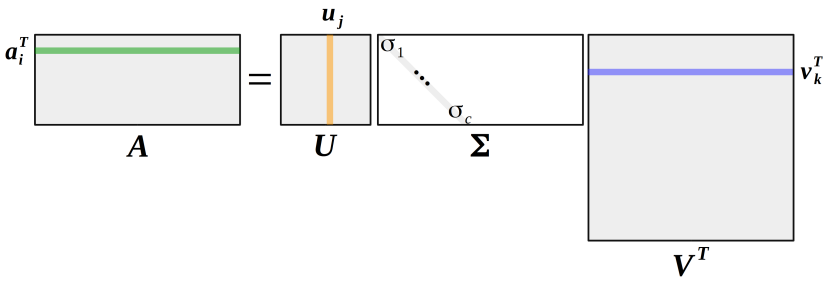


Figure 3: The full singular value decomposition of $\mathbf{A}$ in eqn. **??**.

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

$$\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{V}_r^T, \tag{2}$$

where now $\mathbf{\Sigma}_r$ is a diagonal $r \times r$ matrix that contains only the r non-zero singular values of $\mathbf{A}$ down its diagonal in a non-increasing order, $\mathbf{U}_r = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_r]$ is a $c \times r$ matrix, and $\mathbf{V}_r = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_r]$ is a $g^3 \times r$ matrix. Writing eqn. **??** using an outer product expansion expresses $\mathbf{A}$ as a sum of rank-one matrices:

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{3}$$

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

We numerically compute the singular value decomposition using the `svdfact` function in Julia.[?]

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:

$$\mathbf{x}_s = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_{g^3} \mathbf{v}_{g^3}, \tag{4}$$

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:

$$\mathbf{A}\mathbf{x}_s = \alpha_1 \sigma_1 \mathbf{u}_1 + \alpha_2 \sigma_2 \mathbf{u}_2 + \cdots + \alpha_r \sigma_r \mathbf{u}_r. \tag{5}$$

This follows from $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ for $i = 1, 2, ..., r$ and $\mathbf{A}\mathbf{v}_i = \mathbf{0}$ for $i = r+1, ..., g^3$. Eqn. **??** describes an r-dimensional ellipsoid lying in $\mathbb{R}^c$, whose principal semi-axes are in the direction of $\mathbf{u}_i$ with lengths of σ_i for $i = 1, ..., r$. 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 lays in $\mathbb{R}^c$. The SVD recovers the directions of the principal semi-axes 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 vectors $\{\mathbf{u}_i\}$ are orthonormal and, remarkably, so are 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:

$$\mathbf{A}_\nu = \sum_{i=1}^{\nu} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{6}$$

Comparing to eqn. **??**, the optimal rank ν 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 $\mathbf{A}$ by collapsing the shortest principal axes of the ellipsoid in eqn. **??**; justified intuitively, an ellipse e.g. in 2D is best-approximated by its longest principal axis. The relative error in the approximation is:

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

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. **??** with a ν-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 $\mathbf{A}$ encapsulates the 3D cage cavity images of all 74 porous cage molecules. We showed that we can approximate the data matrix $\mathbf{A}$ with a lower-rank approximant $\mathbf{A}_\nu$ in eqn. **??**. The right singular vectors $\{\mathbf{v}_i\}$ lie in the space of all 3D cage cavity images. As only the first ν right singular vectors appear in the approximant in eqn. **??**, the best ν-dimensional subspace of 3D cage cavity images is thus spanned by the orthonormal set of vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\nu$. Analogous to eigenfaces,[? ?] we declare this set of vectors, which are fictitious 3D cage cavity images [well, with $\bar{\mathbf{c}}$ subtracted and normalized to have magnitude unity] discovered by SVD, the *eigencages*. Algebraically, eqn **??** approximates the flattened 3D cage cavity image of cage k as:

$$\mathbf{c}_k^T \approx \bar{\mathbf{c}}^T + \sum_{i=1}^{\nu} \sigma_i \mathbf{u}_i[k] \mathbf{v}_i^T, \tag{8}$$

confirming cage k is approximately a linear combination of the eigencages $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\nu$ with weights composed of the kth row of $\mathbf{U}_\nu$ and the singular values $\sigma_1, \sigma_2, ..., \sigma_\nu$. In this linear combination, the singular value σ_i appears as a weight to its corresponding eigencage $\mathbf{v}_i$, indicating a hierarchy of importance of the eigencages and justifying discarding the singular vectors with smaller singular values in the approximant in eqn. **??**. Instead of a g^3-dimensional 3D cage cavity image, each porous organic cage molecule can be represented by its composition of $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\nu$ given in eqn. **??**. The *latent representation* of the 3D cage cavity image of cage k is therefore row k of $\mathbf{U}_\nu \mathbf{\Sigma}_\nu$. See Fig. **??**.

We choose the dimension of the latent space ν as the smallest such that the relative error in eqn. **??** is less than 15%, leading to $\nu = 23$. See Fig. **??**. As we have $c = 74$ cages, $\nu = c = 74$ would exactly reconstruct all cages in our cage data set. Thus, this compression to $\nu = 23$ dimensions is a 70% compression of the cage cavity images while incurring only a 15% error in reconstructing the cages. The distribution of the singular values of $\mathbf{A}$ is shown in Fig. **??**.

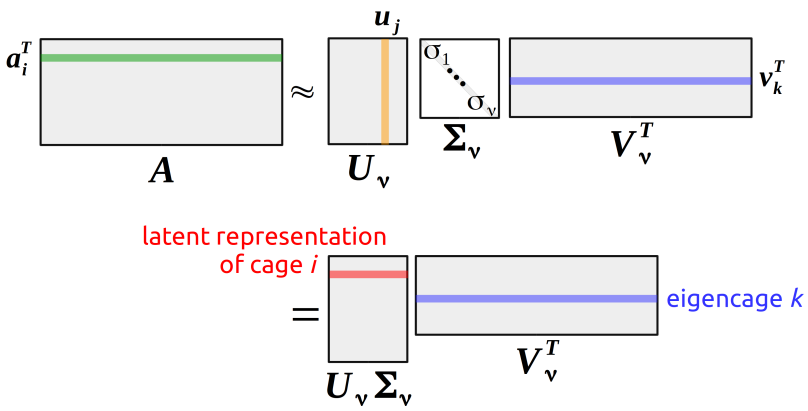


Figure 4: The low rank approximant $\mathbf{A}_\nu \approx \mathbf{A}$ in eqn. **??**. Compare to Fig. **??**. Eigencage k is row k of $\mathbf{V}_\nu^T$. The latent representation of cage i is row i of $\mathbf{U}_\nu\mathbf{\Sigma}_\nu$.

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 cavities of porous organic cage molecules lay. This subspace is spanned by the set of orthonormal right singular vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_\nu$, 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 approximately construct a 3D cage cavity image from a linear combination of the eigencages. The ν-dimensional *latent representation* of cage i is the ith row of $\mathbf{U}_\nu\mathbf{\Sigma}_\nu$.

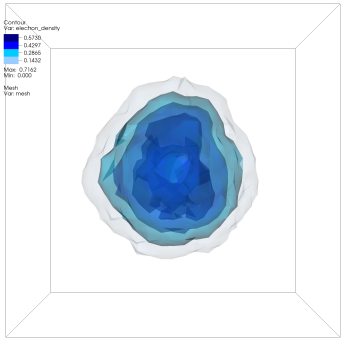
The eigencages

The eigencages– the rows of $\mathbf{V}_\nu^T$ (see Fig. **??**)– are an orthonormal basis for the approximate lower dimensional subspace in which all 3D cage cavity images lay and are ordered in terms of importance. The eigencages are the directions in 3D cage cavity image space that account for the most variance among the 3D cage cavity images in the dataset.[?] We visualize the

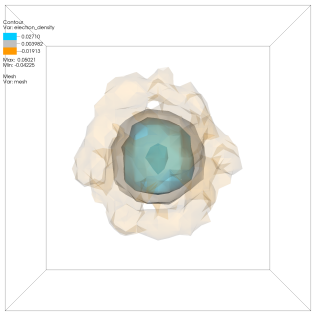
first (and most important) six eigencages in Fig. **??**. As eqn. **??** illustrates, the eigencages express deviations from the average 3D cage cavity image $\bar{\mathbf{c}}$, whose contours are shown in Fig. **??**. Thus, the eigencages possess contours at both positive and negative values. The first eigencage $\mathbf{v}_1$ in Fig. **??** possesses a radially symmetric core; as Fig. **??** indicates, $\mathbf{v}_1$ is used heavily to describe how the cavity diameter of a given cage differs from the average cage in Fig. **??**. 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 eigencage is difficult to reconcile with our intuition of how to describe a cage cavity, 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

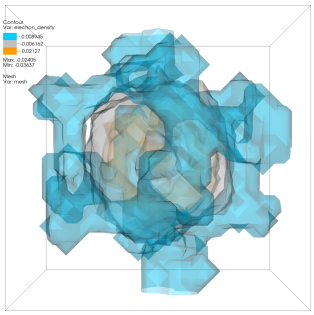
From the view of SVD, each 3D cavity image is constructed from a linear combination of the eigencages (see eqn. **??**). We show here that, remarkably, even the six eigencages in Fig. **??** can be enough to visually reproduce the structure of the cavity in a porous organic cage molecule. As an example, in Fig. **??** we show the approximated reconstruction of the 3D cage cavity image of porous organic cage molecule B25 using different numbers of eigencages. Mathematically, Fig. **??** shows the approximant given by eqn. **??** with $\nu = 1, 2, ..., 9$. Expressing B25 in terms of the first two eigencages is insufficient to capture its shape; for $\nu = 3$, clearly the cavity of B25 is captured to some extent. Only until we reach $\nu = 6$ does the outer-contour of the reconstructed 3D cavity image of B25 well-approximate the shape of the molecule. The reconstructed B25 3D cavity image with $\nu = 9$ is approaching visual indistinguishably from the exact 3D cavity image in Fig. **??**. This shows that e.g. a $\nu = 9$-dimensional latent representation of B25 is sufficient to visually reproduce the shape of its cavity.



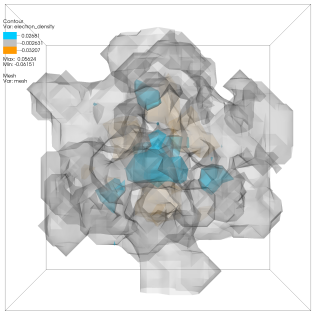
(a) $\bar{\mathbf{c}}$



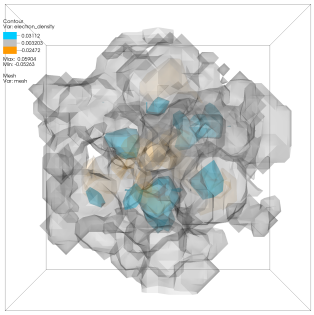
(b) $\mathbf{v}_1$



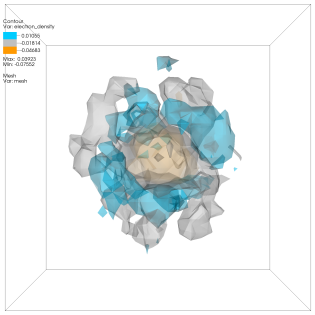
(c) $\mathbf{v}_2$



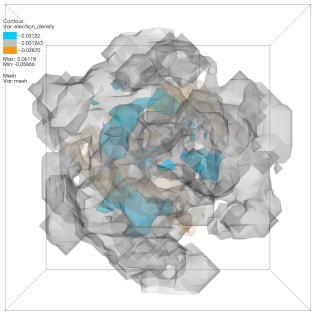
(d) $\mathbf{v}_3$



(e) $\mathbf{v}_4$

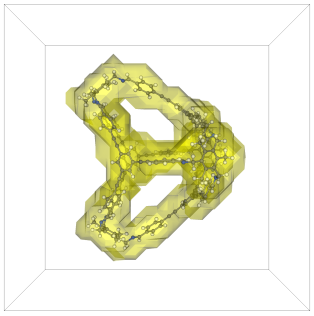


(f) $\mathbf{v}_5$

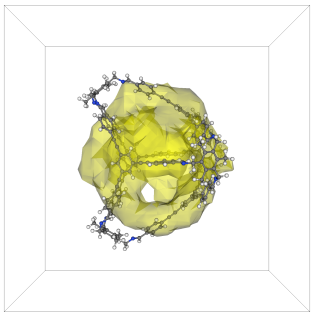


(g) $\mathbf{v}_6$

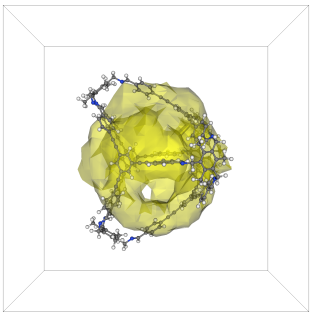
Figure 5: Visualizing the eigencages. (a) Contour surfaces of the average 3D cage cavity image. (b-g) Contour surfaces of the first six eigencages. Orange: low (negative), Gray: intermediate, blue: high (positive).



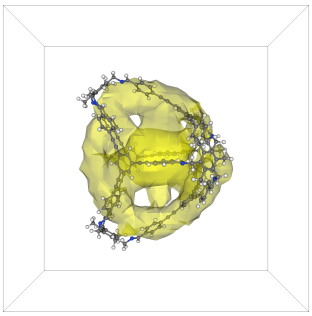
(a) B25, exact $\nu = 74$



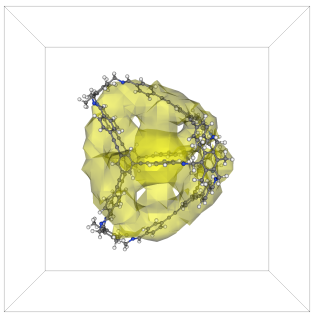
(b) $\nu = 1$



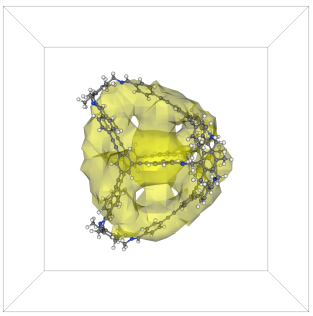
(c) $\nu = 2$



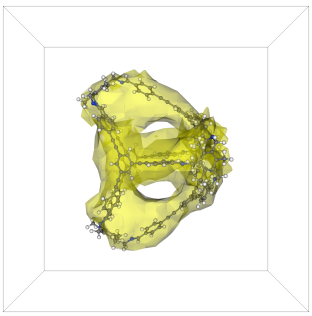
(d) $\nu = 3$



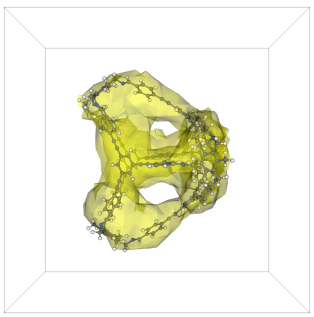
(e) $\nu = 4$



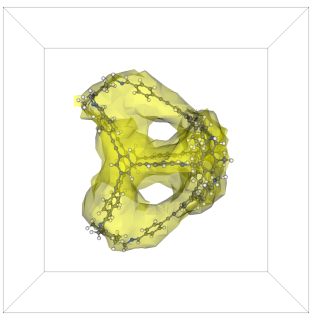
(f) $\nu = 5$



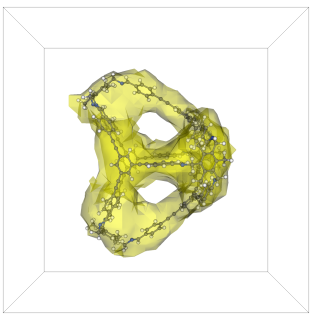
(g) $\nu = 6$



(h) $\nu = 7$



(i) $\nu = 8$



(j) $\nu = 9$

Figure 6: Reconstructing cage B25 with its latent representation. (a) Exact 3D cavity image of B25. (b-j) Reconstructions using latent representations of varying dimensions ν. These are contours (0.5) of eqn. **??** with varying ν.

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. **??**). While the latent space is too large to visualize directly ($\nu = 23$), we resort to t-Distributed Stochastic Neighbor Embedding (t-SNE)[? ?] (perplexity set to 5) to embed our discovered latent space of porous cages into a two-dimensional space for a scatter plot. t-SNE is a non-linear dimensionality reduction algorithm designed to preserve the structure of the data in the embedding. Fig. **??** shows the resulting two-dimensional embedding of the 3D cavity images of the porous cages; each cage appears as a point in this *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. Fig. **??** highlights a few salient clusters. The remaining clusters are shown in Fig. **??**. 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 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.[‡] Compared to human-engineered/invented features of cages, we expect that our discovered latent space defined in $\mathbf{U}_\nu\mathbf{\Sigma}_\nu$ efficiently encodes the salient features of the cavities in porous

[‡]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".[?]

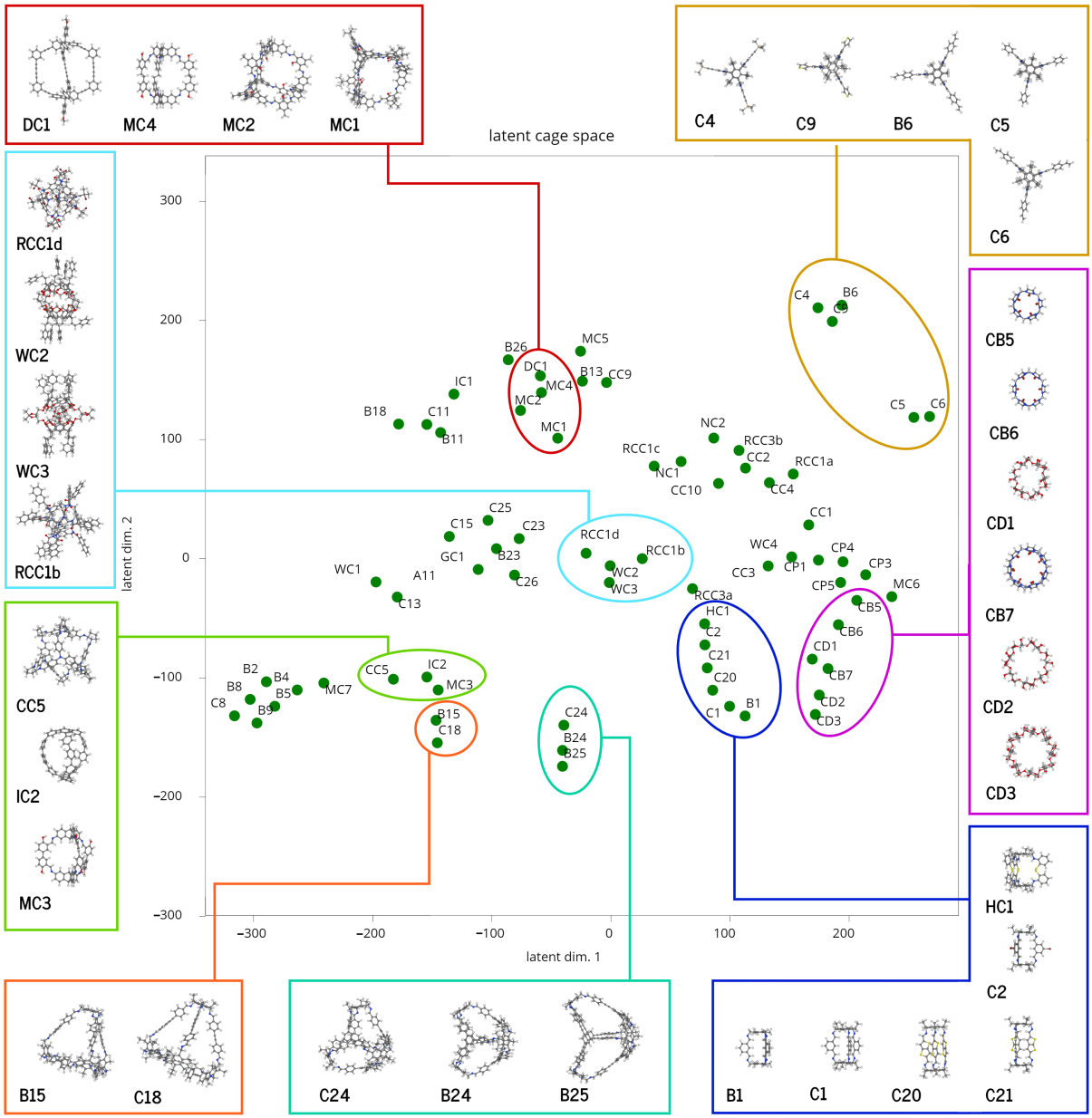


Figure 7: The latent space of cages $\mathbf{U}_\nu\mathbf{\Sigma}_\nu$ embedded into 2D by t-SNE.[??] Salient clusters are highlighted. See Fig. **??** for the remaining clusters.

organic cages. The first reason is that we obtained the latent representations by optimally compressing the 3D cage cavity images with SVD. The second reason is that, as Fig. **??** shows, the latent representation of a cage can visually reproduce the shape of its cavity. The third reason is our observation that 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.[?] See Sec. **??** for details, discussion, and comparison to experimental Xe and Kr adsorption measurements in noria (NC2)[?] and CC3.[?] The points in Fig. **??** are colored according to the simulated Xe/Kr selectivity at infinite dilution. 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. 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. **??** clearly shows that cages nearby in latent space tend to have similar molecule and cavity diameters. Note that MC6, the largest cage in the database, appears to be an outlier. Finally, Fig. **??** shows that cages within clusters in latent space strongly tend to have the same number of windows in which gas molecules can enter the cavity. Together, Figs. **??**, **??**, and **??** 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 separations[? ?]) and thus lends a meaningful notion of similarity between two cage molecules.

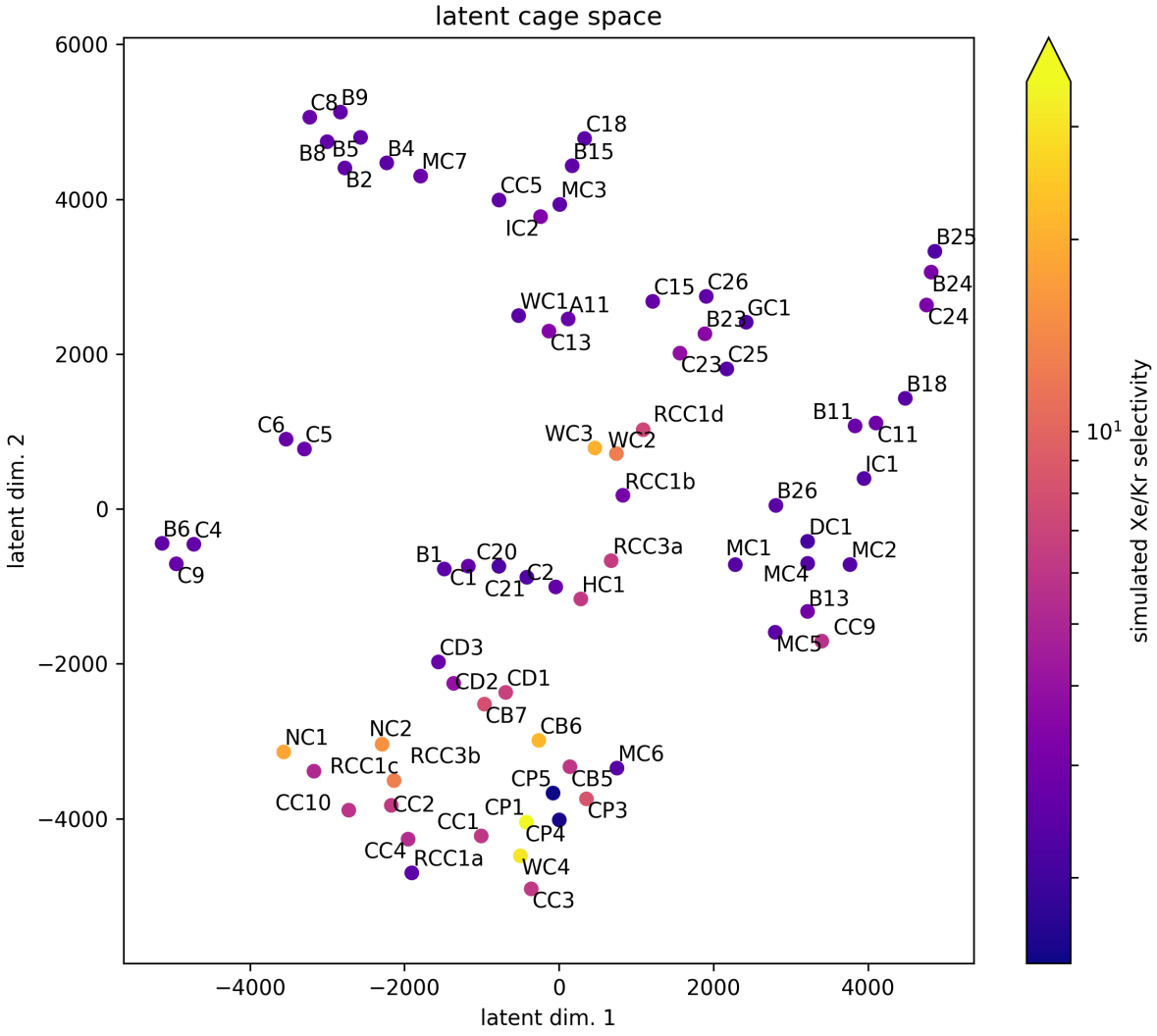
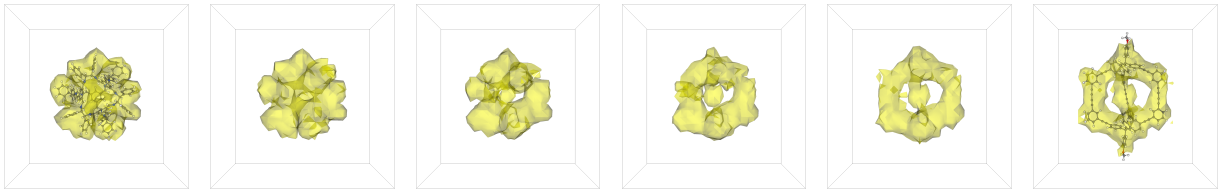


Figure 8: The latent space of cages $\mathbf{U}_\nu\mathbf{\Sigma}_\nu$ embedded into 2D by t-SNE.[? ?] The color of points represents the simulated Xe/Kr selectivity of an isolated cage molecule in an empty box at 298 K. Points nearby in the latent cage space are likely to exhibit similar Xe/Kr selectivities.

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 **??** that we can interpolate between two given cages in the $\nu = 23$ 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.



(a) CC9 (b) (c) (d) (e) (f) DC1

Figure 9: Walking through latent space from cage CC9 (a) to DC1 (f). (b-e) are fictitious cage cavities generated by walking along a line between the latent representation of CC9 and DC1.

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.[?] Since, advanced classes of materials harboring nano-sized pores such as porous cage solids[? ?] have emerged and offer high adsorptive selectivities.[? ? ? ? ? ?] 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 scanned the cavities of 74 porous organic cage molecules to generate three-dimensional images. 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 cage cavity 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 a *latent space* of the cages, 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 $\nu = 23$-dimensional latent representation efficiently encodes the salient features of the cavities in the porous cages displayed in Fig. **??**.

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 cage cavity images by the singular value decomposition sheds some light on the diversity of cavity shapes among cages that have been synthesized. A $\nu = 23$ dimensional latent representation– a 70% compression of the 3D cage cavity images– incurred less than 15% relative error. Consequently, one might suggest that the cavities of the 74 cages in Fig. **??** 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. Considering the 3D cavity images of each cage in its average conformation, however, still may likely lend a useful latent representation. 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![?] That said, our latent representation of the cage molecule geometry could also be predictive of how porous organic cages molecules assemble to form the solid. A third limitation is that the singular value decomposition is sensitive to alignments and translations for inter-cage comparisons; a pair of intuitively-similar cages that are highly asymmetric about the center may not be aligned as intuition would dictate using our alignment approach of diagonalizing the moment of inertia matrix. Finally, 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. 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.

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 lower dimensional subspace. 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 cage cavity image is periodic with varying Bravais lattices. We are exploring

how more advanced dimensionality reduction algorithms that offer translational invariance such as autoencoders[?] with convolutional layers[? ?] may fare; still, rotational invariance is a concern.

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.

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 cage cavity images and Henry coefficients were calculated with our open source code `PorousMaterials.jl` v0.1.1.[?] A Supporting Information document is also available.

Graphical TOC Entry