What are the best materials to separate a xenon/krypton mixture?

Cory M. Simon,[†] Rocio Mercado,[‡] Sondre K. Schnell,[¶,†] Berend Smit,[†,§] and Maciej Haranczyk[*,‖]

University of California, Berkeley, Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Department of Chemistry, Norwegian University of Science and Technology, Trondheim, Norway, Institut des Sciences et Ingenierie Chimiques, Ecole Polytechnique Federale de Lausanne (EPFL), Rue de l'Industrie 17, CH-1950 Sion, and Lawrence Berkeley National Laboratory, Scientific Computing Group

E-mail: mharanczyk@lbl.gov

[*] To whom correspondence should be addressed

[†] University of California, Berkeley, Department of Chemical and Biomolecular Engineering

[‡] University of California, Berkeley, Department of Chemistry

[¶] Norwegian University of Science and Technology

[§] Institut des Sciences et Ingenierie Chimiques, Ecole Polytechnique Federale de Lausanne (EPFL), Valais, Switzerland

[‖] Lawrence Berkeley National Laboratory, Scientific Computing Group

Abstract

Accelerating progress in the discovery and deployment of advanced nanoporous materials relies on chemical insight and structure property relationships for rational design. Because of the complexity of this problem, trial-and-error is heavily involved in the laboratory today. A cost-effective route to aid experimental materials discovery is to construct structure models of nanoporous materials *in silico* and use molecular simulations to rapidly test them and elucidate data-driven guidelines for rational design. For example, highly-tunable nanoporous materials have shown promise as adsorbents for separating an industrially relevant gaseous mixture of xenon and krypton. In this work, we characterize, screen, and analyze the Nanoporous Materials Genome, a database of ca. 670,000 porous material structures, for candidate adsorbents for xenon/krypton separations. For over half a million structures, the computational resources required for a brute-force screening using grand-canonical Monte Carlo simulations of Xe/Kr adsorption are prohibitive. To overcome the computational cost, we used a hybrid approach combining machine learning algorithms (random forests) with molecular simulations. We compared the results from our large-scale screening with simple pore models to rationalize the strong link between pore size and selectivity. With this insight, we then analyzed the anatomy of the binding sites of the most selective materials. These binding sites can be constructed from tubes, pockets, rings, or cages and are often composed of non-discrete chemical fragments. The complexity of these binding sites emphasizes the importance of high-throughput computational screenings to discover new materials. Interestingly, our screening study predicts that the two most selective materials in the database are an aluminophosphate zeolite analogue and a calcium based coordination network, both of which have already been synthesized but not yet tested for Xe/Kr separations.

Introduction

Nanoporous materials have applications in gas storage,[1] gas separation,[2] gas sensing,[3] catalysis,[4] and drug delivery.[5] By combining different molecular building blocks in their synthesis, advanced classes of nanoporous materials are highly tunable. For example, in metal organic frameworks (MOFs),[6] metal nodes or clusters form a coordination network with organic linkers. Other highly adjustable materials include covalent organic frameworks (COFs),[7] zeolitic imidizolate frameworks (ZIFs),[8] and porous polymer networks (PPNs).[9] High chemical tunability enables one to tailor-make a material for each application under a variety of conditions, but also inundates researchers with practically endless possibilities. Due to limited resources and time in practice, only a small subset of the possible materials can be synthesized and tested for each application. A component of the Materials Genome Initiative[10,11] is to use computational tools to navigate this vast chemical space[12] of materials and rapidly test them *in silico* to accelerate the discovery and deployment of new materials and aid experimental efforts for clean energy, security, and human welfare.

Within the Nanoporous Materials Genome,[13] we explore the chemical space of possible nanoporous materials by constructing materials *in silico*; this is achieved by combining different molecular building blocks in various topologies, much like snapping together Lego blocks or K'NEX. In the Nanoporous Materials Genome Center,[13] over 670,000 structures have been constructed to date: libraries of predicted MOFs,[14] all-silica zeolites,[15,16] PPNs,[17] ZIFs,[18] and COFs[19] as well as data sets of zeolites[20] and MOFs[21] that have been experimentally synthesized. Henceforth, we will refer to this set of materials as the *Nanoporous Materials Genome*. We can rapidly and cost-effectively test each of these materials for a given application using mathematical models and molecular simulations.

This type of computational *screening* focuses experimental efforts on the most promising materials, elucidates trends or guidelines for synthesizing an optimal material, and identifies performance limits. High-throughput screenings of porous materials can be found in the literature for vehicular natural gas storage,[22] carbon capture,[18,23] hydrogen storage,[24] ethene/ethane separations,[25] ethanol purification,[26] and sulfur dioxide removal.[27]

In this work, we screen the libraries of materials in the Nanoporous Materials Genome for high-performing materials for room temperature xenon/krypton separations. The noble gases xenon (Xe) and krypton (Kr) have several important applications.[28] Xenon is used as an anesthetic[29–31] and for imaging[32] in the health industry and as a satellite propellant in the space industry.[33] Both xenon and krypton are used in lighting,[34] in lasers,[35,36] in double glazing for insulation,[37,38] and as carrier gases in analytical chemistry.[39] As krypton and xenon are present in Earth's atmosphere at concentrations of 1.14 and 0.087 ppm, respectively,[40] the conventional method to obtain xenon and krypton is as a byproduct of the separation of air into oxygen and nitrogen by cryogenic distillation.[41] This byproduct stream from air separation consists of 80% krypton and 20% xenon.[42] At Air Liquide, this mixture is compressed to 200 bar and stored in cylinders, then sent to a separate Xe-Kr separation plant to undergo *another* cryogenic distillation to obtain pure xenon or krypton.[43] Cryogenic distillation for the separation of krypton and xenon has a very high energy and capital requirement, reflected by the cost of high-purity xenon, ca. $5,000/kg.[44] An alternative, potentially energy- and cost-saving Xe/Kr separation process is to use a nanoporous material as an adsorbent to selectively adsorb either Kr or Xe at ambient temperature through a temperature- and/or pressure-swing adsorption process.[45,46] Xenon and krypton are also products of the nuclear fission of uranium and plutonium;[40] porous materials could be used to capture the radioactive xenon and krypton in the processing of used nuclear fuel.[47–49] Experiments regarding xenon and krypton adsorption[44,48,50–62] suggest that it may be feasible to use nanoporous materials in an adsorption-based process to separate a Xe/Kr mixture.

For an adsorption-based separation of xenon and krypton, we seek to exploit the differences in their size and van der Waals interactions with the framework atoms of the nanoporous materials. Xenon is larger than krypton (van der Waals radii: $r_{Xe} = 1.985$ Å and $r_{Kr} = 1.83$ Å) and has a deeper Lennard-Jones attractive potential well (see Figure S1). Thus, we expect the majority of materials to selectively adsorb xenon over krypton; other materials will be reverse selective if their pores are the appropriate size for krypton, but too small for xenon.

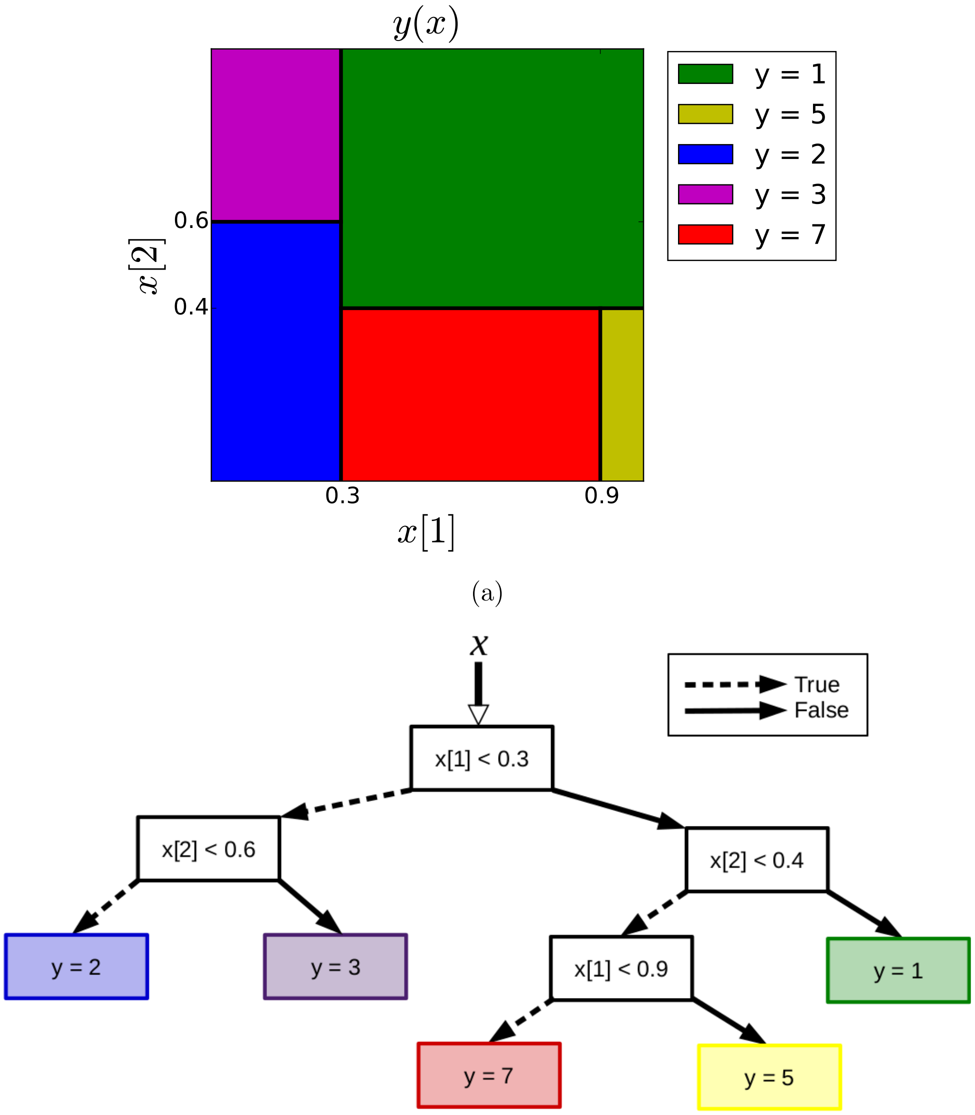
A brute-force high-throughput screening for Xe/Kr separations has been done on the hypothetical MOF dataset[43,63] and a set of experimental MOFs.[64] We extend these studies to include PPNs, ZIFs, zeolites, and COFs. With such a rapid increase in the number of materials, the required computational resources will become a bottleneck for these brute-force screening techniques. Therefore, it is important to develop algorithms that allow us to screen these databases more efficiently. In this work, we utilize machine learning algorithms to predict performance from structural descriptors. Simple structural descriptors, such as the surface area and void fraction, can be quickly computed. Some structural descriptors, such as crystal density and pore size, are independent of the application and conditions studied and thus can be computed once, stored as a material property, and utilized in several different contexts. These structural descriptors are often correlated to and thus predictive of material performance, underlying the concept behind quantitative structure-property relationship (QSPR)[65] modeling. As such, the array of structural descriptors of a material serves as a representation/fingerprint of the material in a high-dimensional space, a *feature vector*. Machine learning techniques, such as random forests,[66] support vector machines,[67] and neural networks,[68] can then be trained to predict material performance from the feature vector (for a review, see Ref.[65]). With this pre-screening technique, we focus the molecular simulations on only a fraction of the total number of structures, the most promising ones, and avert simulating adsorption in materials that are most likely poor performers.

In the realm of nanoporous materials, Fernandez and coworkers[69] showed that machine learning techniques can predict methane uptake in MOFs from geometric descriptors. In another work, Fernandez and coworkers[70] illustrated the classification of MOFs as high- or low- performing for CO_2 uptake using support vector machines. These works illustrated machine learning as a potential method to be utilized in high-throughput screenings of materials.

We utilize a random forest,[66] an ensemble[71] of decision trees, to predict the Xe/Kr separation performance from structural descriptors. For a given material, we predict the selectivity by running the corresponding vector of descriptors through a set of decision trees. Figure 1 illustrates a decision tree regressor. A decision tree is grown using a training set of materials with known selectivities. However, if we train a single decision tree on the entire training set, the prediction is not very accurate, as a single decision tree is highly prone to overfitting.[72] Overfitting leads to an accurate representation of the training set but a high variance prediction for materials that are not in the training set, thus yielding poor generalization error.[72] Much better accuracy on unseen data can be obtained by training many trees – a forest of trees – on randomly selected subsets of the training set, decorrelating the trees from each other. The prediction of the selectivity in a *random forest* is then the average vote among each tree in the forest. The reason that an ensemble of decorrelated trees gives a greater accuracy than a single tree, a Wisdom of the crowds[73] effect, is as follows. Each tree is allowed to overfit its own training data. Thus each tree in the forest accurately describes its own training data, and so we retain a low bias prediction, i.e., the average does not systematically shift away from the true average. While each tree has a high variance prediction on unseen data on its own, by taking the average of each tree in the forest, we reduce the variance, leading to a low-bias, low-variance prediction. Consequently, random forests have been shown to be remarkably accurate[74] in a wide variety of problems. For example, in computer vision, by training a random forest on a set of pictures of humans in different positions, Microsoft successfully implemented random forests to accurately label body parts from video in real-time for the Kinect in the XBox video game console.[75] Similarly, by feeding a random forest training examples of porous material structures and their Xe/Kr separation performance as predicted by molecular simulation, we show that random forests can successfully expedite the discovery of high-performance porous materials.

The success of machine learning in screening materials is predicated on having highly-predictive structural descriptors. In this work, we introduce a new descriptor, the Voronoi energy E_v, that takes into account both geometrical structural information and the energetics of the guest-host interaction. The Voronoi energy is the average energy of a xenon atom at the accessible Voronoi nodes that represent the pore topology; we find this descriptor to be highly predictive of Xe/Kr separation performance.

With the large volume of data from our hybrid machine learning- molecular simulation screening approach, we uncover guidelines for synthesizing a highly selective material. In particular, we rationalize the strong relationship between performance and pore size by considering two simple caricatures of spherical and cylindrical adsorption pockets. With this insight, we then analyze in detail the top candidates from the Nanoporous Materials Genome for Xe/Kr separations. We identify the characteristic chemistry that makes a binding site selective for these type of separations. This molecular insight provides some guidelines for a possible rational design of these materials.



(a)

(b)

Figure 1: Decision tree regressor illustration for the function $y(x)$, where x is in two-dimensions. (a) A decision tree regressor partitions the feature space into rectangles and models $y(x)$ as a constant in each one. The plane shows the two-dimensional feature space, and the color indicates the value $y(x)$ assigned to a feature vector x that belongs to that region. (b) The decision tree shown takes a feature vector x at the root node (top) and represents $y(x)$ depicted in (a). At each node, a binary decision is made by comparing a component of x and a threshold. The data point is sent to the left (right) if condition in the box is true (false). The terminal/leaf nodes assign a value of y. The colors of the leaf nodes correspond to the regions represented in (a).

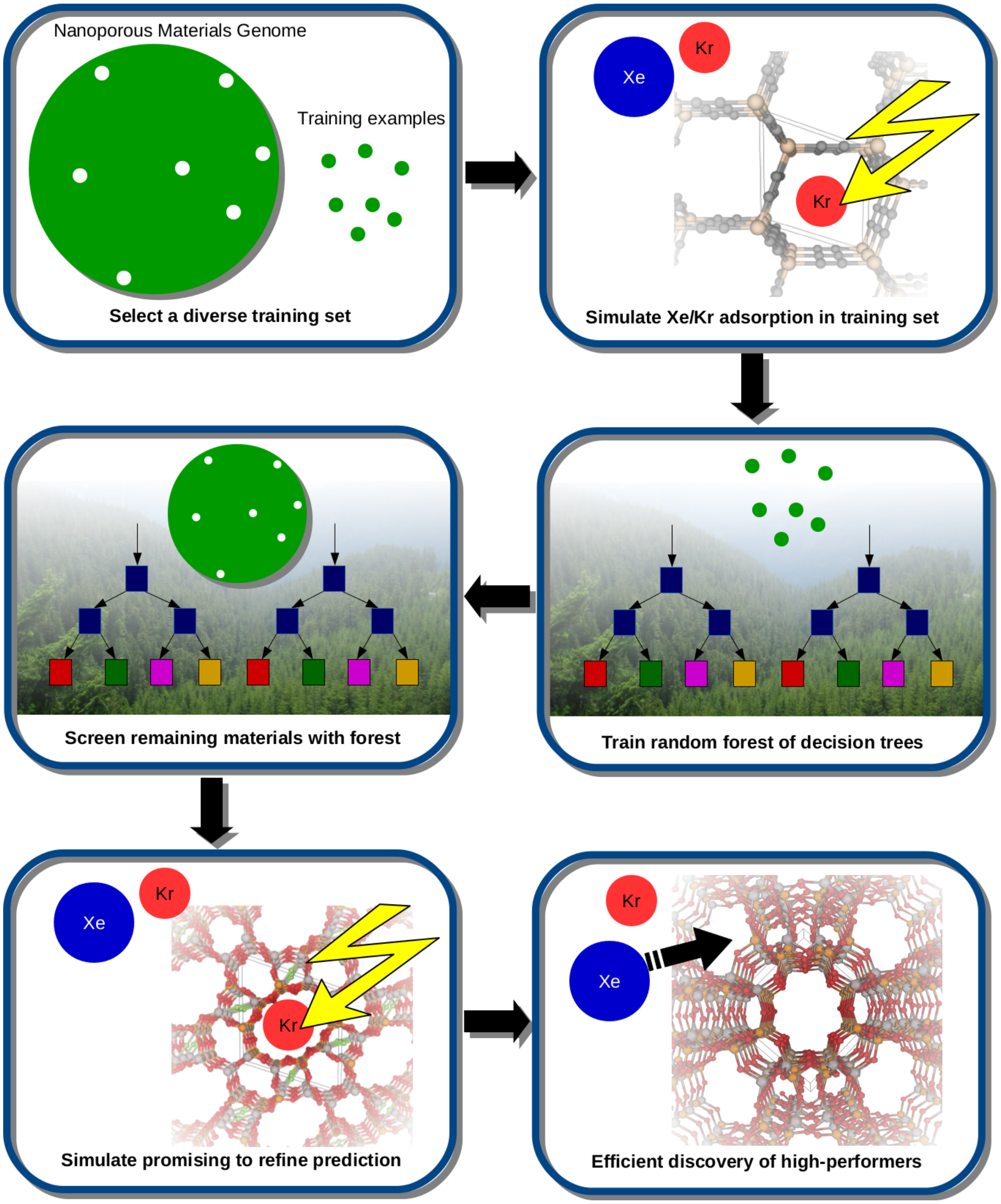

Nanoporous Materials Genome
Training examples
Select a diverse training set
Xe
Kr
Kr
Simulate Xe/Kr adsorption in training set
Train random forest of decision trees
Screen remaining materials with forest
Xe
Kr
Kr
Simulate promising to refine prediction
Kr
Xe
Efficient discovery of high-performers

Figure 2: Screening procedure. We first select a mathematically diverse training set from the structures in the Nanoporous Materials Genome and simulate the adsorption of Xe/Kr in this training set. Then, we train a random forest of decision tree regressors to predict performance from structural descriptors. We run the structural descriptors of the remaining materials through the trained random forest to identify the most promising materials. We simulate Xe/Kr adsorption in the most promising set to refine the prediction. Note that some high performing materials are discovered during training.

Methods

The workflow of our method to screen ca. 670,000 structures in the Nanoporous Materials Genome is illustrated in Figure 2 and consists of the following six steps. 1. Characterization: we compute the feature vector of each material, whose components are seven quickly-computed structural descriptors. 2. Selection of the training set: we use a diversity selection algorithm[76] to ensure that our training set of materials adequately covers our seven-dimensional feature space. 3. Label the training set: we perform binary grand-canonical Monte Carlo simulations to compute the Xe/Kr uptake in the training set. 4. Training of the forest: we use these training examples to grow a forest of decision tree regressors to predict Xe/Kr separation performance from the structural descriptors. 5. Pre-screening: we run the structural descriptors of the remaining materials through the trained forest of decision tree regressors. 6. Materials discovery: if the forest predicts the material to be promising for Xe/Kr separations, we run grand-canonical Monte Carlo simulations to refine the prediction.

Molecular simulations

The potential energy of interaction of an adsorbate (Xe or Kr) with the atoms of the material and other adsorbates is modeled with pairwise Lennard-Jones potentials. The Lennard-Jones parameters for Kr-Kr and Xe-Xe interactions are taken from Ref.[77] The parameters for MOF, ZIF, COF, and PPN atoms are taken from the Universal Force Field.[78] For zeolites, the Lennard-Jones parameters for oxygen are back-calculated using Lorenz-Berthelot mixing rules from a force field for krypton developed specially for zeolites by Talu and Myers;[79] in this force field, Kr-Si interactions are not explicitly considered and are instead embedded in the Kr-O parameters. The cross-interaction Lennard-Jones parameters of Xe/Kr with atoms of the framework are then calculated from the Lorenz-Berthelot mixing rules.

We truncate the Lennard-Jones potentials at 12.5 Å and approximate interactions beyond the cutoff distance to be zero. We implement periodic boundary conditions to mimic an infinite crystal and assume the structures to be rigid, which induces some errors since some structures are flexible upon gas adsorption.[80,81]

We simulate the equilibrium Xe and Kr uptake in a material with the grand-canonical Monte Carlo algorithm[82] for a two-component system. We consider a 20/80 molar Xe/Kr mixture at 298 K and 1 atm of total pressure. The Monte Carlo moves are particle exchange (insertion/deletion), particle translation, and particle identity change with probabilities 0.6, 0.2, and 0.2, respectively. We utilize a parallel algorithm written for graphics processing units (GPUs)[83] to pre-compute potential energy interpolation grids (ca. 0.1 Å grid point spacing) to speed-up our simulations.

As evidence that our force field can reasonably predict the experimental Xe/Kr uptake in materials, simulated adsorption isotherms of Xe and Kr match well with the experimental Xe/Kr uptake in metal-organic frameworks IRMOF-1 and IRMOF-2x, for $x =$Cl, Br, F, I,[53] and Silicalite.[84] With a generic force field such as the UFF, it is particularly difficult to predict uptake in MOFs with open metal sites.[85] Still, for the purposes of a high-throughput screening, where we seek to identify the top candidates, the UFF provides a similar *ranking* of materials compared to highly accurate but more computationally expensive *ab initio* force fields.[86] See Sec S2.3 for detailed comparisons of simulated and experimental Xe/Kr adsorption isotherms.

Some pores in nanoporous materials are large enough to fit an adsorbate molecule, but are inaccessible in a periodic system due to prohibitively high energy barriers to diffuse into the pore. In our molecular simulations, we identify such pockets automatically in a high-throughput manner and refrain from sampling adsorbate configurations in these regions using an inaccessible pocket-blocking algorithm in Ref.[87] (see Sec S2.1).

Structural descriptors

Table 1 summarizes the structural descriptors used to form the feature vector of the material. The first six descriptors are established geometric structural descriptors. The last descriptor is, to our knowledge, a new descriptor that we invented that incorporates both geometrical structure information and guest-host interaction energy.

While one can imagine many other structural descriptors, those in Table 1 were easily computed and, as we will later see, are satisfyingly predictive of Xe/Kr separation performance for our screening.

Table 1: Structural descriptors used to construct a feature vector for each material.

Descriptor	Symbol	Description
Void fraction [unitless]	ϵ_v	fraction of material that is free volume
Crystal density [kg/m^3]	ρ	mass of crystalline material per volume
Largest free sphere diameter [Å]	D_f	largest sphere to percolate through material
Largest included sphere diameter [Å]	D_i	largest sphere to fit inside the material
Accessible surface area [m^2/cm^3]	a	probe-accessible surface area along pore walls
Surface density [kg/m^2]	ρ_s	mass of atoms per acccessible surface area $= \rho/a$
Voronoi energy [kJ/mol]	E_v	average energy of Xe at the accessible Voronoi nodes

Established geometric structural descriptors

To compute geometric structure descriptors, we use open-source software Zeo++,[88] which models the accessible void space space inside a porous material with a periodic Voronoi network. The framework atoms as well as Xe/Kr are modeled as hard spheres with radii adopted from the Cambridge Structural Database[89,90] and the zeros of the Lennard-Jones potentials in Figure S1, respectively. The periodic Voronoi network is calculated by Voro++.[91] Figure 3 shows a two-dimensional sketch of the Voronoi network in a toy material. From this periodic graph representation of the pore space, we can identify regions that are accessible to Xe and Kr and calculate the diameter of the largest included and free spheres (D_i and D_f),[92] accessible surface area (a), and void fraction (ϵ_v).[88] The crystal density (ρ) is easily computed from the crystal structure. Another descriptor we use is the surface density ρ_s, i.e. the mass of material per surface area, $\rho_s = \rho/a$ [kg/m^2].

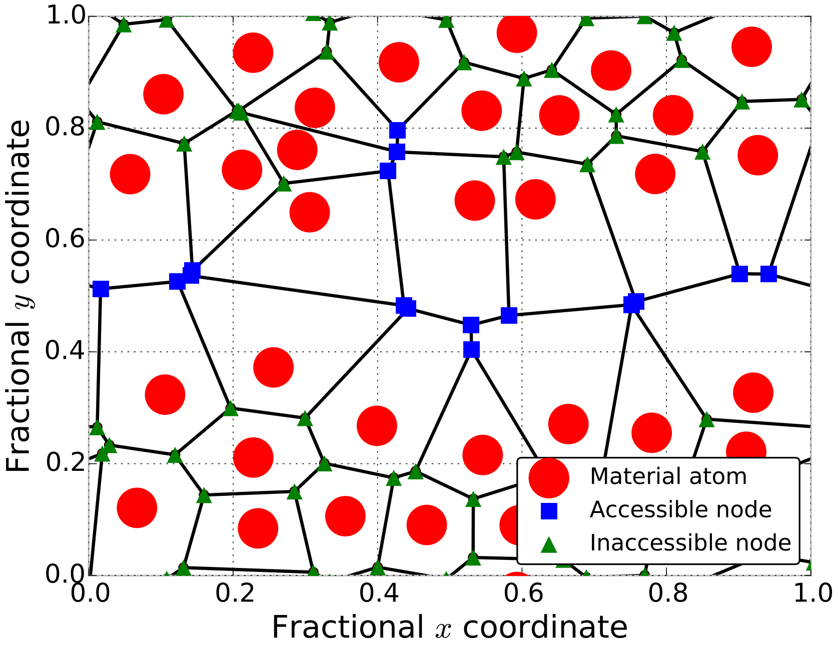


Figure 3: Voronoi network model of void space (2D caricature). The unit cell of a toy material is shown. Red circles represent atoms of the material; accessible and inaccessible Voronoi nodes are blue squares and green triangles, respectively. The black lines are the edges in the periodic Voronoi graph that models the void space. The descriptor E_v in eq 1 is the average potential energy of a xenon atom adsorbed at the accessible Voronoi nodes.

The Voronoi energy descriptor

Gas uptake at low pressures can be dominated by the strongest adsorption sites due to the exponential contribution of the energy in the Henry coefficient.[82] The established geometric structural descriptors above have difficulty capturing such “sweet spots”.[93] In an attempt to capture such strong binding pockets that are important for Xe/Kr adsorption, we included a descriptor E_v [kJ/mol] that is the average energy of a xenon atom among the nodes of the xenon-accessible Voronoi network:

$$E_v := \frac{1}{N} \sum_{i=0}^{N} E(v_i), \tag{1}$$

where $E(v)$ is the potential energy of a xenon atom at a point v in the pore space of the material as calculated by our force field. The set of points $\{v_i\}$ for $i = 1, 2, ..., N$ are the accessible Voronoi nodes in the Voronoi network used to model the xenon-accessible pore space; the Voronoi network conveniently and automatically identifies points that are located at the ‘center’ of the accessible pore landscapes (see Figure 3). Thus, the descriptor E_v can be viewed as a biased sampling of the potential energy of an adsorbate in the pores. In a material with pores that are near the optimal size for xenon, the energetic minimum will be at the center of the pores, where Xe can maximally interact with all surrounding atoms (see Figure 7(d) later). As the Voronoi energy E_v is a measure of energy at the ‘center’ of the pore, our intuition is that E_v will capture these strong binding regions in materials with nearly optimal pore diameters. Local structural features, such as a pocket, can also be identified automatically with the Voronoi network, for example, the pocket in Figure 3. In contrast, the energetic minimum in materials with large pores will be near the surface instead of the center; however, materials with large pores are poor for Xe/Kr separations (see Figure 5 later), and D_i should take responsibility to rule out materials with large pores in the random forest algorithm. Later, we will find that our new descriptor E_v is highly predictive of separation performance.

Selecting a diverse set of materials

To ensure that the training data set adequately represents the chemical space[12] of crystal structures in our dataset and for the decision forest that we train to be accurate, the training data set must include a *diverse* set of materials. To pose a mathematical definition of diversity, we represent each material in a high dimensional space with a *feature vector* x and utilize the Euclidean distance as a notion of similarity. For diversity selection, we define the feature space *a priori* as $x = [\rho, a, D_i, \epsilon_v]$ and normalize each component of the vector to range from zero to unity. To select the diverse set, we use an algorithm in Ref.,[76] which we outline in Sec S2.4. We selected a diverse subset of ca. 15,000 materials to serve as training data (3,640 experimental MOFs, 2,750 hMOFs, 1,000 hPPNs, 2,319 hZeolites, 194 IZA zeolites, 3,757 hZIFs, and 834 hCOFs ['h' indicates hypothetical or predicted materials that have not been synthesized]).

Forest of decision tree regressors

With a random forest of decision tree regressors, we can quickly predict the Xe/Kr separation performance from the seven-dimensional feature vector x_i of a material. We will characterize the performance of a material by the selectivity in eq 2, but we will denote the performance score here as y_i for a general discussion. Using the training data set (x_i, y_i) for $i = 1, 2, ..., N$, we show here how to train a forest of decision tree regressors that serves as our regression $y(x)$.

A decision tree[94] resembles a flowchart and assigns a value $y(x)$ to a feature vector x by the path it takes through the tree. Geometrically, a single decision tree regressor partitions feature space into a set of hyperrectangles and models the output y as a constant in each partition.[68] For an illustration, Figure 1 shows a decision tree and corresponding function $y(x)$ for the case of a two-dimensional feature vector.

To train/grow a decision tree, we consider splits at each node of the form $x[i] < k$. The component of the feature vector (i) and threshold (k) are chosen using a greedy algorithm.[68] See Section S2.5 for a more formal description. We used open-source scikit-learn[95] in Python to implement the decision forest. For our trees in the random forest, we grow each tree to the maximum extent possible.

A strong advantage of decision tree regressors is interpretability; we can quantify the importance of a particular feature in improving the quality of prediction by keeping track of the reduction of the mean square error at each node and to which feature this reduction in error is attributed.[94] We will later plot the relative importance of each feature in our random forest. Other advantages of decision trees are: (1) normalization or scaling of the data is not required, as decision trees are invariant to monotone transformations of the features, (2) irrelevant features will not severely detriment performance, as the decision nodes can simply ignore these unimportant features and not consider them for splits,[68] and (3) decision trees are relatively robust to outliers.[72] Still, a [single] decision tree is often not as accurate as other models, in part due to the greedy construction procedure, and is highly prone to overfitting.[72]

A *decision forest*[66] is an ensemble[71] of randomly trained decision trees. In this ensemble method, by taking the average vote among a committee of diverse/de-correlated decision trees,[71] we obtain much greater accuracy and generalization than by using a single decision tree,[96] a Wisdom of the Crowds[73] effect. We inject diversity into the trees of the forest by inducing randomness during training using the following techniques: (1) bagging (bootstrap aggregating), where each tree is trained on a random subset of the training data,[97] (2) random split selection,[98] where at each split node, the split is selected at random from the K best splits, and (3) the random subspace method,[99] where a random selection of the features are considered for each tree. The prediction of the forest is the average prediction among each tree in the forest.

In general, the utility of a model for making predictions is determined by its expected error over future data not seen during training time – the generalization error.[72] Typically, one sets aside a validation data set or performs cross validation.[68,72] As we use bagging in our random forest, despite training on the full data set, we essentially set aside a validation set during training in the process of leaving out a subset of the data for each tree in the forest during the bagging procedure. After training each tree, we take the data left out from the training during bagging, run it through the tree, and store the prediction. In this way, we get the *out of bag error*, which is the error of the forest's prediction on data unseen during training[68] and a quality metric of how well our forest will perform on future data.

Results

We consider a crystalline nanoporous material in equilibrium with a 20/80 molar Xe/Kr mixture at 298 K and 1 bar. A good material for Xe/Kr separations should have a high selectivity $S_{Xe/Kr}$ for Xe over Kr:

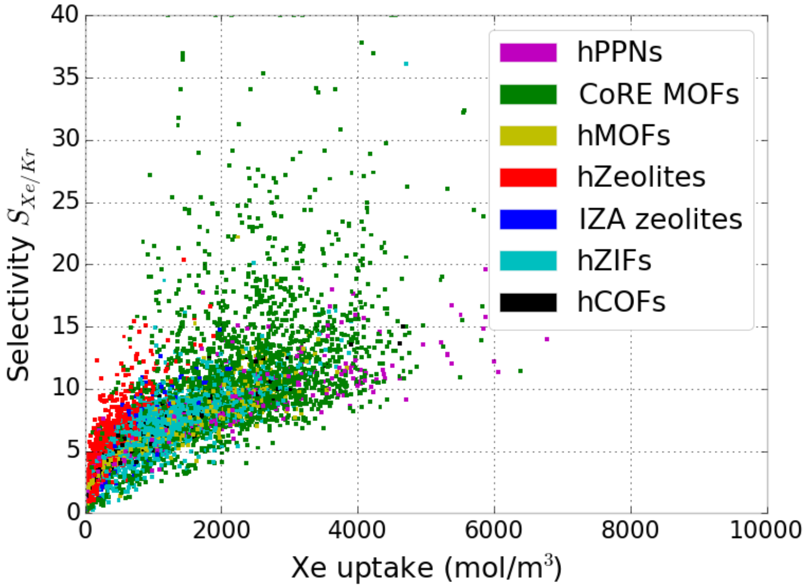
$$S_{Xe/Kr} := \frac{(x_{Xe}/x_{Kr})}{(y_{Xe}/y_{Kr})}, \tag{2}$$

where $x_{\{Xe,Kr\}}$ is the mole fraction in the adsorbed phase and $y_{\{Xe,Kr\}}$ is the mole fraction in the bulk gas phase ($y_{Xe} = 0.2$, $y_{Kr} = 0.8$ in our study). A good material could also be reverse selective if its pores are too small for and thus exclude Xe but are large enough to accommodate Kr. In addition to a high Xe selectivity, a good material should also have a high Xe uptake to reduce the amount of material that is needed and the size of the adsorption column to obtain a given amount of Xe.

Training data

Using the diversity selection algorithm and $x = [\rho, a, D_i, \epsilon_v]$ as a feature vector, we selected a mathematically diverse training set of ca. 15,000 materials. The feature space explored by these materials and a scatter plot matrix of these features are depicted in Figure S3 and Figure S4. We then performed grand-canonical Monte Carlo simulations to predict the uptake of xenon and krypton in each material in the training set (20/80 molar Xe/Kr mixture at 298 K and 1 atm). The key performance plot for Xe/Kr separations in Figure 4(a) shows the simulated Xe/Kr selectivity against the Xe uptake in the training set (as in Ref.[43]); the best materials lie in the upper-right region.

For machine learning algorithms to be a successful pre-screening tool, the structural descriptors must be correlated to the Xe/Kr separation performance. Scatter plots of selectivity against our structural descriptors indeed show correlations (Figure S9). As an example, Figure 5 shows a strong relationship between the selectivity and the largest included sphere; the highest selectivities occur when the pore diameter is close to the distance that yields the optimal energy in the Xe-Xe Lennard-Jones potential ($2^{1/6}\sigma_{Xe-Xe}$), plotted as a vertical, dashed line (more on this later). Despite the pore size playing an important role and having some predictive value, other variables are at play; there are several structures with this optimal pore size that have poor selectivities. There is not a simple recipe of a given geometric structural feature that guarantees a material with a high selectivity (Figure S9). This motivates the utility of our random forest of decision tree regressors that predict performance by “looking” in higher dimensional feature space and simultaneously considering all structural features. The exceptionally predictive structural descriptor is the Voronoi energy.



(a)

Figure 4: Performance plot regarding training data selected from the diversity selection algorithm. Xe/Kr selectivity $S_{Xe/Kr}$ against volumetric Xe uptake. Each structure is a point, and color indicates material class. In this plot as well as others, points are plotted in a random order to avoid seeing only data points plotted last. In Figure S11, we show the performance plot for each class of material separately. The performance plot using gravimetric xenon uptake closely resembles this because volumetric and gravimetric xenon uptake are correlated (see Section S8).

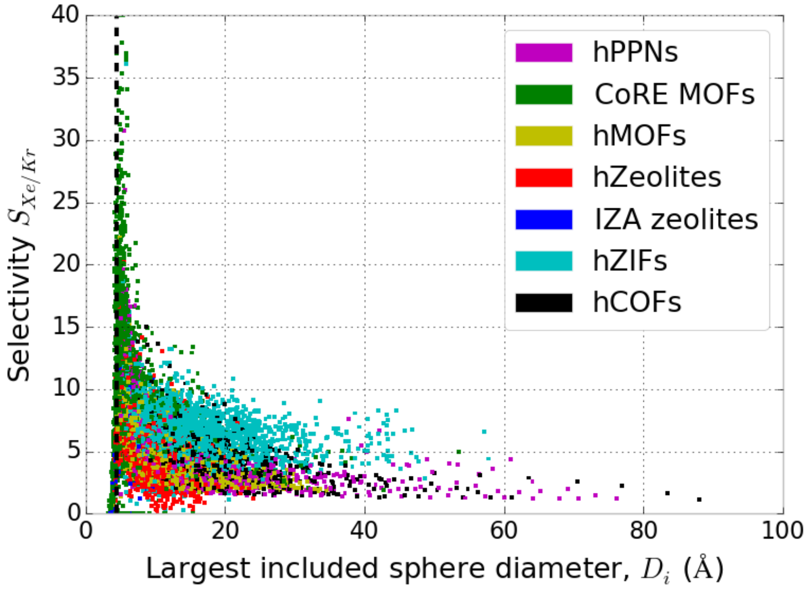


Figure 5: Selectivity against pore size. Xe/Kr selectivity $S_{Xe/Kr}$ against diameter of largest included sphere D_i. Vertical, dashed line is $2^{1/6}\sigma_{Xe-Xe}$, the optimal Xe-Xe distance. Selectivities clipped at 40.

Training the random forest

In the absence of a purity specification and an economic model of a Xe/Kr separation plant (describing capital costs of equipment and material and recurring costs of regenerating the material) to construct a sensible performance metric of a material, we use the selectivity for the performance score y of a material for parsimony. This score implicitly takes xenon loading into account since selectivity positively correlates with Xe uptake in Figure 4(a). For the decision forest to predict the selectivity of a material, we use as a feature vector $x = [\rho, \epsilon_v, a, D_i, D_f, E_v, \rho_s]$ (see Table 1 for descriptions).

We then grew a random forest of 1000 decision tree regressors using the diverse training set. Figure S12 shows that 1000 is an adequate number of trees for our regression. We excluded materials with inaccessible pockets as determined by our pocket blocking algorithm (see Methods) to avoid confusing the trees, as a subset of the geometric structural descriptors cannot possibly predict this phenomena that nonetheless could significantly affect the simulated selectivity. In total, we trained the random forest using 9,376 training examples. We used four features in the random subspace method,[99] as this yielded the smallest out of bag error.

Figure 6(a) displays a histogram of the simulated selectivity $S_{Xe/Kr}$ in the training set of materials against the out of bag prediction (see Methods) by the random forest. The red, diagonal line indicates where the random forest's prediction exactly matches the simulated selectivity; the root mean squared error is 1.2. We observe a high variance in the predictions of the decision forest for materials with the highest selectivities, possibly due to the scarcity of data in this performance range, which we will later address, or the inadequacy of our descriptors for these materials (irreducible error).

An interesting question is what features are most important in predicting selectivity. We plot the relative importance of each feature (see Methods) during training, averaged over each tree in the forest, in Figure 6(b); a higher feature importance implies the feature played a larger role in decreasing the mean square error in the decision tree regressor during training. We see that the expected energy of a xenon atom at the accessible Voronoi nodes, E_v, is by far the most important feature. This is not surprising, given that energy is a major thermodynamic driving force for adsorption. The second most important feature is the void fraction. We duly note that correlations between variables can influence this feature importance plot.[66,100] For reference, the correlations between the features in our training set are depicted in the correlation matrix in Figure S5. To further support that the Voronoi energy E_v is the most predictive descriptor, we performed feature subset selection; in both forward selection and backward elimination,[68] E_v is the first feature selected and last feature eliminated, respectively. An exhaustive search over all possible combinations of features in the random forest regression showed that, in selecting the top d predictors, E_v was included in the best feature combination for every d. For details, see Section S5.

Screening the remaining materials

The good agreement between the out of bag prediction and simulated selectivity in Figure 6(a) indicates that the trained random forest can reasonably predict the selectivity in the remaining (outside of the training set) ca. 655,000 materials for which we did not yet perform molecular simulations. To screen the remaining materials, we run their feature vectors of structural descriptors through the trained random forest. If the prediction $\hat{S}_{Xe/Kr} > 11$ (to the right of the green vertical line in Figure 6), we refine the prediction by a grand-canonical Monte Carlo simulation of the Xe/Kr adsorption. In this manner, we focus the computational resources for molecular simulations on the materials predicted to be the most selective by the random forest. We choose a threshold selectivity of 11 in part because the random forest exhibits low variance in the region $\hat{S}_{Xe/Kr} \leq 11$, and thus we have greater confidence in its prediction.

The trained random forest predicted that 4,066 remaining structures satisfy the condition $\hat{S}_{Xe/Kr} > 11$, which we refer to as the *promising set*. In Figure 6(c), we show the simulated selectivities of the promising set against the prediction by the random forest. The root mean squared error is 2.2. This includes materials with blocked pockets that the decision trees were not trained to predict, which account for most of the outliers in Figure 6(c).

We visualize the efficiency of our screening strategy by comparing the distribution of simulated selectivites in the promising set with the distribution in a *randomly* selected set of ca. 1000 materials from the Nanoporous Materials Genome in Figure 6(d). The mode of the simulated selectivity distribution for the promising set is significantly greater than that for the randomly selected set, confirming the efficiency of our screening strategy using random forests. The random forest selects the best materials from the database and circumvents wasting simulation time on poorly performing materials, as in a brute-force screening strategy. For comparison, we also show the distribution of selectivities in the diverse training set, which mimics that of the random set.

It is important to ensure that the random forest does not incorrectly label promising materials as low performing. The match between simulation and out of bag prediction for materials with low selectivities in Figure 6(a) is evidence already. To test this further, we simulated Xe/Kr adsorption for the 500 materials in the remaining set that the random forest predicted to have the lowest selectivities. These simulations confirmed (see Section S14) that indeed these materials have low selectivities.

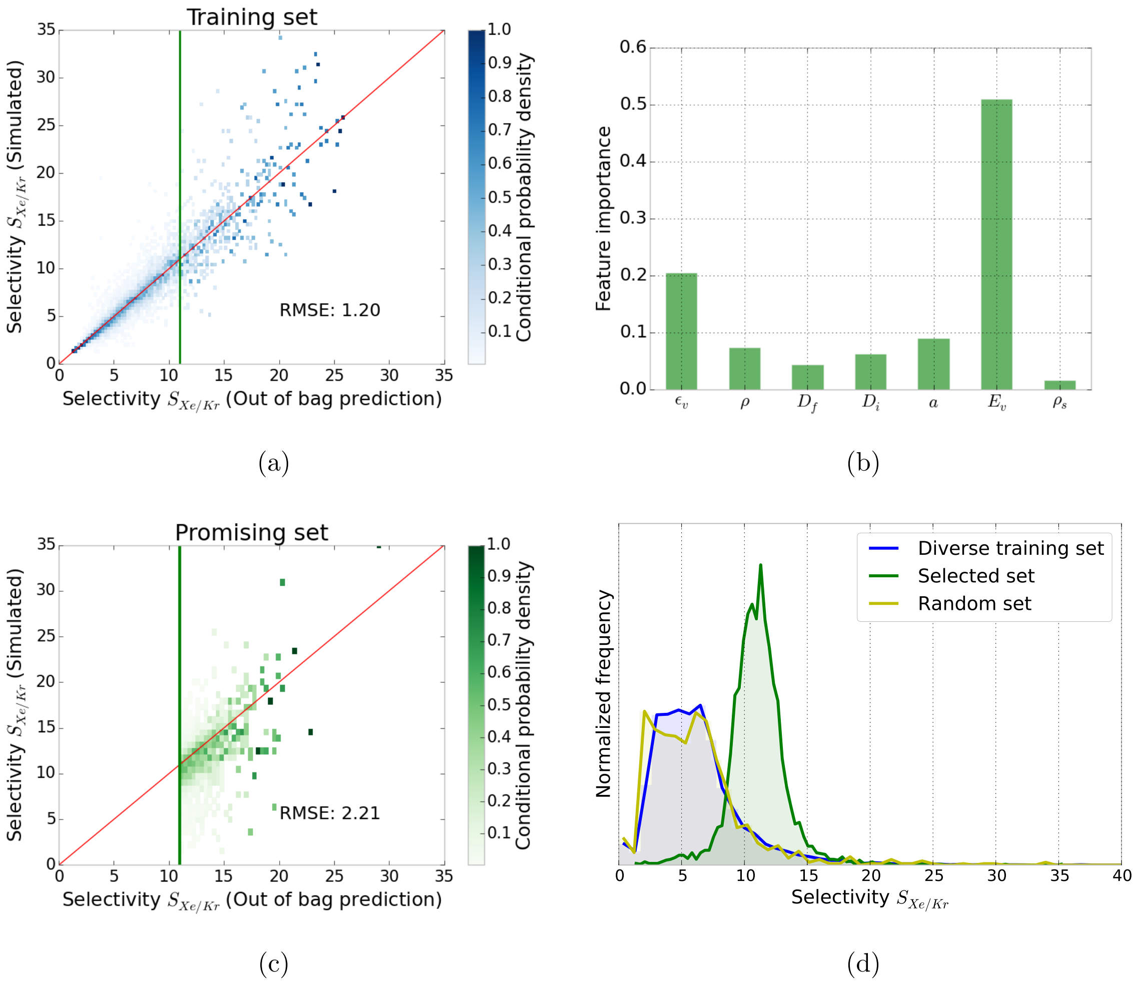
To ensure that the decision forest is not under-predicting the selectivity of some materials because of the paucity of example structures with a high selectivity in the diverse training set (Figure 6(a)), we trained a decision forest using the *union* of the diverse training set in Figure 6(a) and the promising set in Figure 6(d), providing more training examples of high-performing materials. We then ran the feature vectors of the remaining materials through this re-trained random forest and find 461 additional materials that meet the criteria $\hat{S}_{Xe/Kr} > 11$. We then simulated Xe/Kr adsorption in these materials to refine the prediction. We did not discover any additional materials beyond those found in the first round of training with exceptionally high selectivities (see Figure S13). Thus, we are confident that our training set selected by the diversity selection algorithm adequately explores feature space.

By utilizing random forests, we performed grand-canonical Monte Carlo simulations in only 20,000 of the 670,000 materials in our process of screening the Nanoporous Materials Genome for Xe/Kr separations and utilized the quickly-computed feature of structural descriptors to rule out the rest. Before looking at the top materials from our screening, we first gain insight into how the pore size and shape affect the selectivity.

The role of pore size and shape

Snurr and coworkers found that the highest Xe/Kr selectivities in MOFs were obtained with pores around the size of a Xe atom,[43] as we see in Figure 5. To grasp how the selectivity of a material depends on its pore size, we study here two simple model materials which capture the relationship between Xe/Kr selectivity and pore size that we observe in Figure 5. Since a given diameter of the largest included sphere D_i can be achieved with differently shaped pores, for example, a spherical and cylindrical pore with the same radius, we also investigate the effect of pore *shape* on the selectivity.

Consider two model adsorption pockets, one constructed from a spherical shell of radius R and the other constructed from an infinite cylindrical shell of radius R. See Figure 7(a). The framework atoms are uniformly distributed on the surface of the sphere/cylinder. We model the interactions between Xe/Kr and atoms of the shell using Lennard-Jones potentials (energy parameter ϵ, size parameter σ). By integrating the Lennard-Jones potential over the surface of the sphere and cylinder – "smearing" the atoms over the shell surfaces[101] – we obtain effective interaction potentials between an adsorbate and the shells that are functions of r, the radial coordinate of the guest molecule in the spherical[102]/cylindrical[103] shell:



(a)

(b)

(c)

(d)

Figure 6: Decision forest results. (a) Two dimensional histogram of simulated selectivity against selectivity predicted from feature vector by the random forest. This includes all materials in the training set. The x-axis shows the out of bag prediction; each tree contributing to the prediction for a given structure has not been trained on it. The red, diagonal line is a perfect prediction. For perspective, all materials predicted to lie to the right of the green, vertical line are deemed as promising and assigned to the *promising set*. (b) The average feature importance among the decision tree regressors in the random forest. A higher feature importance implies that the feature is more important in reducing the mean square error. See Table 1 for feature descriptions. (c) Two dimensional histogram of simulated against predicted selectivity for the *promising set* of materials (outside the training set). (d) Distribution of simulated selectivities in the diverse training set, the *promising set*, and a random selection of the materials. To ensure a proper comparison, we fixed the percentage of the random set that consists of each material class to be equal to that of the diverse training set.

$$U_{sphere}(r;R) = \eta 4\epsilon \frac{2\pi R}{r}\left\{\frac{\sigma^{12}}{10}\left[\frac{1}{(R-r)^{10}} - \frac{1}{(R+r)^{10}}\right] - \frac{\sigma^6}{4}\left[\frac{1}{(R-r)^4} - \frac{1}{(R+r)^4}\right]\right\}$$
$$U_{cylinder}(r;R) = \eta 4\epsilon 2\pi\left[\frac{\sigma^{12}}{R^{10}}I_6 - \frac{\sigma^6}{R^4}I_3\right] \quad (3)$$
$$I_n := B\left(n-\frac{1}{2},\frac{1}{2}\right){}_2F_1\left(n-\frac{1}{2},n-\frac{1}{2};1;\frac{r^2}{R^2}\right).$$

Here, $B(x,y)$ and ${}_2F_1(a,b;c;z)$ are the beta and hypergeometric funtions, respectively, and η is the surface density of atoms on the shell. These shell models are useful to intuit the relationship between pore size and Xe/Kr separation performance because they tease out the effects of varying pore shapes, topologies, compositions, and surface density of atoms that are present in Figure 5 for a given pore size.

For parsimony, we perform all calculations in this section at infinite dilution and assume the adsorption isotherms are in the Henry regime so that loading can be approximated as K_HP, where K_H is the Henry coefficient and P is the partial pressure. To compare our model to real material structures, we computed the Kr and Xe Henry coefficients in the dataset of experimental MOFs[21] using Widom insertions.[82] We imposed a uniform, constant surface density of framework atoms, η, and investigated the effect of composition in our model material by using different types of atoms (σ and ϵ in eq 3). We found that Universal Force Field[78] Lennard-Jones parameters for silicon and $\eta = 0.1$ Å^{-2} recapitulate the trends of the Henry coefficients of the experimental MOFs.

Using the potentials in eq 3, we calculate the Henry coefficient of xenon and krypton in the model materials as a function of pore size R via Widom insertions.[82] We plot the selectivity at infinite dilution for various shell sizes in Figure 7(b). In our model, we define the largest included sphere diameter $D_i := 2(R - r_{Si})$ to correspond to that calculated by Zeo++, where $r_{Si} = 2.1$ Å is the van der Waals radius of silicon.[89] The models recapitulate the trend in the experimental MOFs (black points), and we observe a well-defined pore diameter that maximizes the selectivity (sphere: 5.7 Å, cylinder: 4.7 Å). The shell models also recapitulate the performance plot of the experimental MOFs in Figure 7(c). As these crude model materials recapitulate the trends in the experimental MOFs, it is evident that the pore size plays a central role in Xe/Kr separations.

Let us rationalize the poor performance of materials with large R. Figure 7(d) shows the Xe and Kr potentials in a large sphere ($R = 8$Å) and in a smaller sphere that yields maximal potential well depth for Xe ($R = 4.38$Å). For the larger sphere, the minimum potential energy occurs on a *surface* near the walls; adsorbates at the center ($r = 0$) are too far from the walls to feel strong interactions. As R decreases, eventually the minimum energy becomes a *point* at the center $r = 0$, as we see with the smaller sphere. Thus, one reason large pores are ineffective is that the attractive interactions at the center are weak and resemble an empty space. The size of the shell also determines the degree of potential overlap by multiple framework atoms contributing van der Waals interactions to recruit the adsorbate. In the large pore, an adsorbate near the surface sees a relatively flat surface, whereas in the smaller pore, an adsorbate is proximal to the shell in all radial directions. As a result of these combined interactions, the potentials in Figure 7(d) are deeper in the smaller pore. More importantly, the *difference* in energy of interaction between a Xe and Kr is greater in the smaller pore, yielding a higher selectivity. Of course, if the pores get too small, repulsive forces become dominant and the selectivity and adsorption decrease rapidly with decreasing R. These observations hold for the cylinder as well.

Note that the radius R of shell that yields the minimal energy for Xe (sphere: 4.38 Å) is larger than that for krypton (sphere: 4.2 Å). Thus, *two* factors contribute to the selectivity for xenon over krypton in the model materials: (1) xenon has a deeper Lennard-Jones attractive potential well (larger ϵ) than krypton and (2) the pore diameter that maximizes interactions with xenon is larger than that for krypton, a size effect (see Figure S1). We can assess the extent to which the size difference is responsible for the selectivity of the model material for xenon by constructing an artificially enlarged krypton molecule by assigning the Lennard-Jones σ to be equal to that of xenon and but keeping the ϵ equal to that of krypton. By simulating the selectivity of our model material for xenon over this artificially enlarged krypton, the peak selectivity was observed to decrease approximately three-fold (See Section S13.2). Thus, the difference in the sizes of Xe and Kr accounts for 2/3 of the selectivity of our model material, while the difference in Lennard-Jones potential depths accounts for the other third.

Finally, we remark on the differences between the spherical and cylindrical shell. A spherical shell has a deeper potential well than a cylindrical shell with the same R and surface density of atoms (Figure S16). The reason is that a xenon atom in a sphere is encompassed by framework atoms in all directions, whereas inside a cylinder, lateral interactions, from the direction along the cylindrical axis, are lacking. As a result, with the same η, we achieve a higher selectivity with the spherical shell in Figure 7(b). Also note that the R that yields the maximal selectivity is larger in a sphere than in a cylinder. This is largely because a Xe or Kr atom in the cylindrical shell experiences its minimum energy in the (r, R) parameter space at a smaller R than in the spherical model (Figure S17). The sphere has the minimum energy at $R = 2^{1/6}\sigma$, the distance that yields the minimum in the adsorbate interaction with the atoms of the shell. The cylinder has a smaller optimal radius, as the suboptimal interactions with the closest atoms are compensated by increasing more interactions in the lateral/axial direction.

While our discussions have centered around energy, we duly note that entropy plays a role in determining the Henry coefficient as well.[82]

The scatter about the model curves exhibited by the experimental MOFs in Figure 7(b) is due to real materials having more complex shapes than perfect cylinders and spheres as well as varying compositions and surface densities. We did not observe a clear difference in the pore size-selectivity relationship between MOFs with 1D channels and 2D/3D channels (Figure S20) because of these significant and complex perturbations in real materials. While these simple models give us great intuition, the deviation of real materials from spheres and cylinders leads us to perform a detailed analysis of the pores of the top performing materials next.

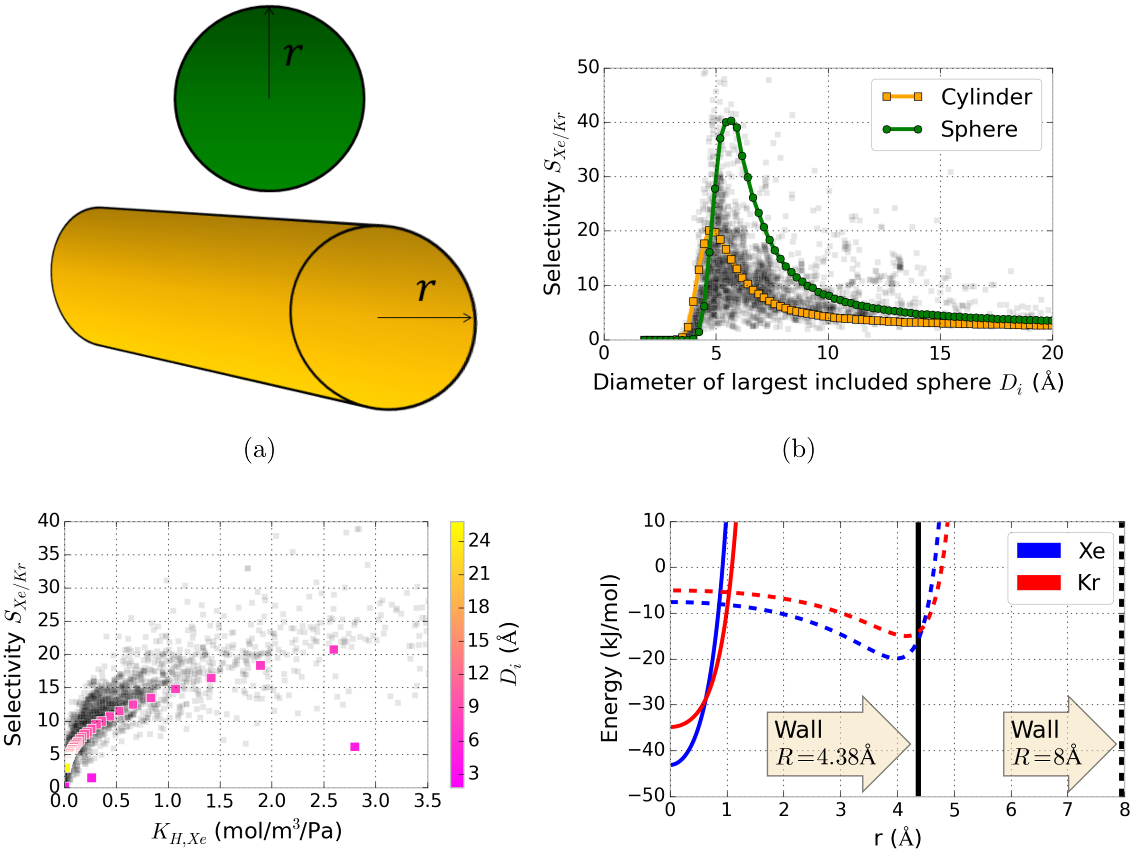
Top performing materials and binding site analysis

We show the performance plot including the *promising set* in Figure S14. Also shown is porous organic cage solid CC3[104] with groundbreaking measured Xe/Kr separation performance,[44] simulated with a force field developed especially for it in Ref.[44] Our models predict many more selective materials for Xe/Kr separations than CC3, which has a selectivity of 13.8 at our conditions. An interactive dashboard with the performance plot data is openly available at:

`http://nanoporousmaterials.org/xekrseparations/`.

Top candidates

We display the structures of the top performing materials in Figure 8. The two structures with the highest selectivity in the Nanoporous Materials Genome are JAVTAC, an aluminophosphate zeolite analogue that has been synthesized by Cooper *et al*,[105] and KAXQIL, a calcium based coordination network synthesized by Banerjee *et al.*[110] Both structures are found in the experimental MOF database.[21] Figures 8(a) and 8(b) show that KAXQIL and JAVTAC consist of 1-dimensional channels. The structure with



(a)

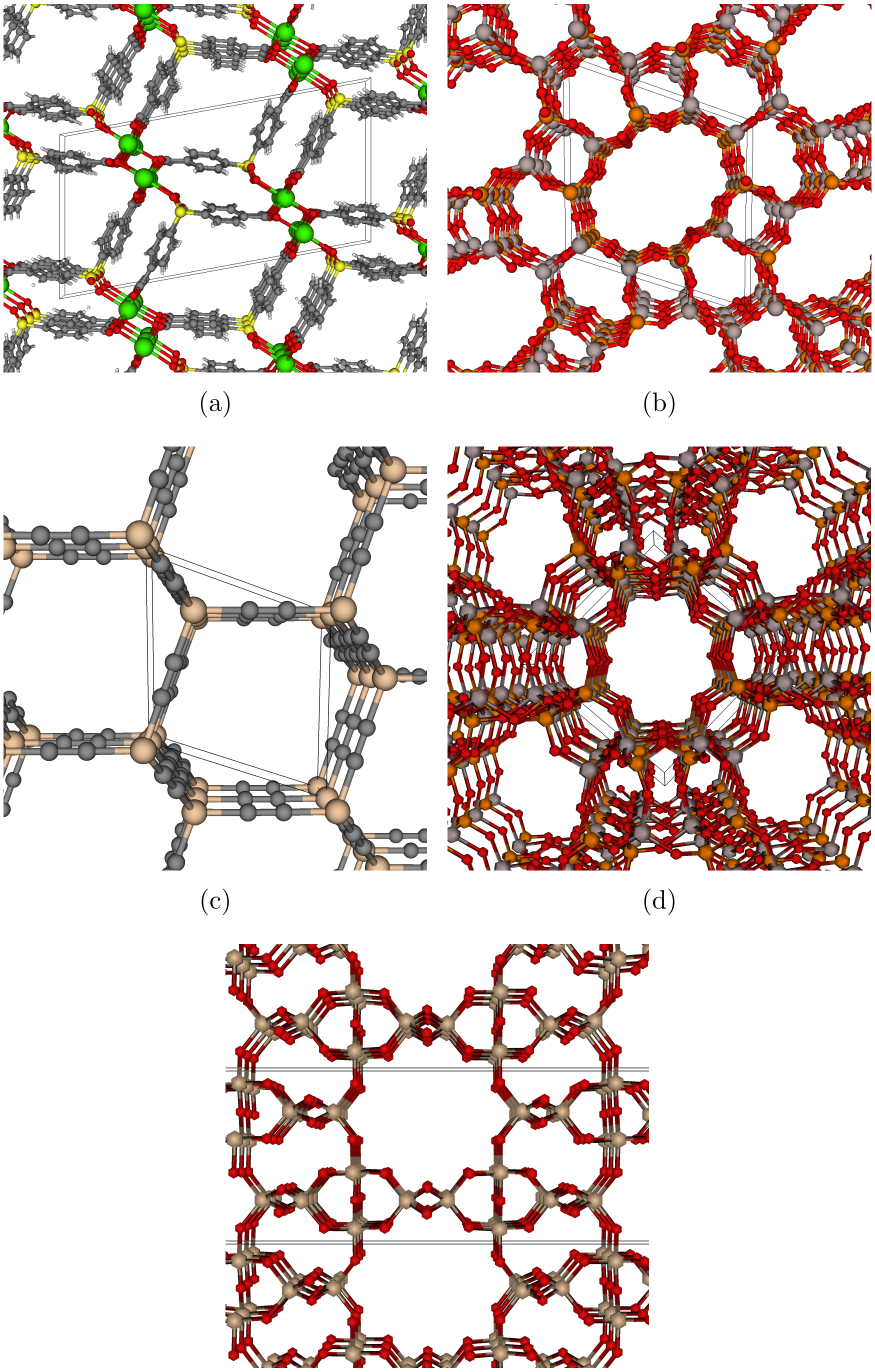
(b)

(c)

(d)

Figure 7: Spherical and cylindrical shell models. (a) In the model, framework atoms are uniformly distributed across the shell and 'smeared' over the surface, leaving an effective potential which is a function of the radial coordinate r. (b) The selectivity in the model materials as a function of the pore diameter. (c) The performance plot of the spherical shell model. Each point is a sphere of a different radius R; colormap indicates D_i. Performance plot for cylinder is similiar (Figure S19). In both (b) and (c), the black, opaque points in the background are the experimental MOFs. (d) The Lennard-Jones potentials in eqn 3 for Xe (blue) and Kr (red) for two differently sized spherical shells. Dashed lines are for the larger sphere ($R = 8$ Å); solid lines are for the smaller sphere ($R = 4.38$ Å).

the highest volumetric Xe loading (8223 mol/m^3) is predicted porous polymer network hPPN_Si_4080_1-net_001,[17] which has three-dimensional channels (see Figure 8(c)). This structure is a diamond analogue, in which connections between C atoms are extended with -C≡C- groups, therefore the name extended diamond (extDIA). This extDIA struc-



(a)

(b)

(c)

(d)

(e)

Figure 8: Top performing structures. (a, b) Calcium coordination network KAXQIL[110] (a, yellow:S, gray:C, red:O, green:Ca) and aluminophosphate zeolite analogue JAVTAC[105] (b, orange:P, gray:Al, red:O) have the highest simulated selectivities ($S_{Xe/Kr}$ =82). (b) Predicted porous polymer network hPPN_Si_4080_1-net_001 has the highest Xe loading. (tan:Si, gray:C) (c) Aluminophosphate molecular sieve GOMREG has the highest product of selectivity and Xe loading. (orange:P, gray:Al, red:O) (d) A reverse selective predicted zeolite h8166173. (tan:Si, red:O)

ture was previously highlighted for its methane storage capability.[106] The structure with the highest product of selectivity and xenon loading is an aluminophosphate molecular sieve[107] that has been synthesized,[108] GOMREG ($S_{Xe/Kr}$ = 76, Xe loading =4500 mol/m^3, also found in the experimental MOF database[21]), shown in Figure 8(d) to consist of 1-dimensional channels.

Some materials are reverse selective because their pores are too small for and thus exclude xenon but are large enough to accommodate a krypton atom. The most reverse selective material we found is a predicted all-silica zeolite[15,16] h8166173, shown in Figure 8(e), with a selectivity $S_{Xe/Kr}$ =0.10. The diameter of the largest included sphere in h8166173 is $D_i = 3.73$ Å, between the van der Waals diameter of a krypton and xenon atom ($2r_{Kr} = 3.66$; $2r_{Xe} = 3.97$ Å).

Characteristics of top materials

To further confirm that there is not a simple formula of structural descriptors to yield a material with a high selectivity, we investigate the difference in structural descriptors between highly and poorly selective materials. Figure 9 compares the distributions of each structural descriptor among highly selective ($S_{Xe/Kr} \geq 14$) and poorly selective ($S_{Xe/Kr} < 14$) materials separately.

Highly selective materials have low void fractions, very low (favorable) xenon energies at the accessible Voronoi nodes (E_v), and pore sizes around that of a xenon atom (reiterating Figure 5). Interestingly, highly selective materials tend to have lower surface areas.

Aside from D_i and D_f, the distributions of geometric structural properties among the most selective materials are very broad; if there were a simple recipe of a structural descriptor to yield an optimal material, the distributions would be narrow. Furthermore,

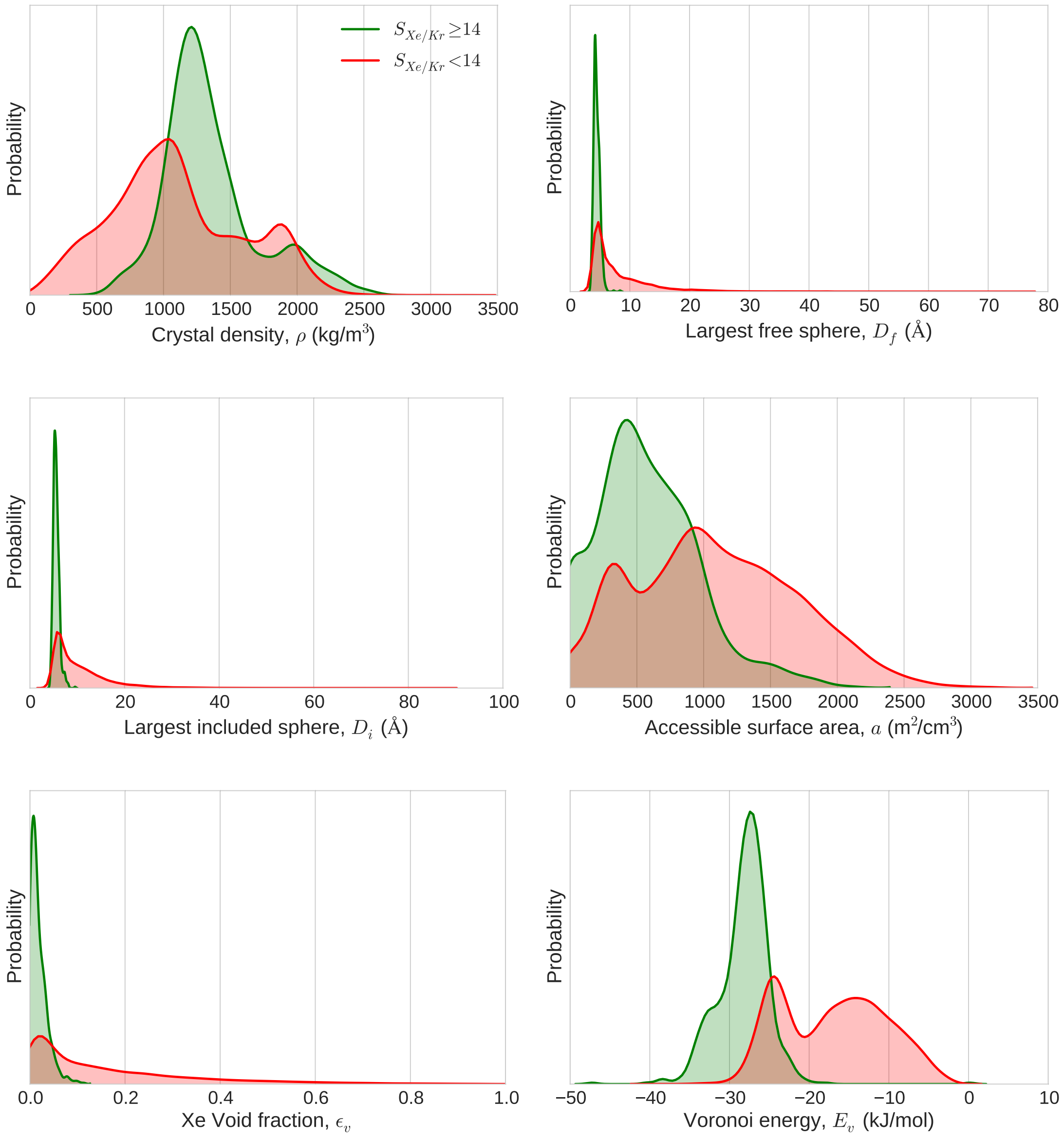


Figure 9: Distributions of structural descriptors explored by highly selective ($S_{Xe/Kr} \geq 14$, green) and poorly selective ($S_{Xe/Kr} < 14$, red) materials separately. The distributions are visualized using kernel density estimation with a Gaussian kernel, with the bandwidth chosen by Scott's rule.[109]

the distributions of geometric properties for the poorly and highly selective materials overlap to a large extent. This again motivates using machine learning to look in this high-dimensional feature space to consider all features at once. The exception is the Voronoi energy descriptor, which, while predictive for screening, is not a straightforward quantity to tune in the synthesis of new materials.

Sikora et al.[43] found that the most selective hMOFs for xenon have cylinder-like pores in contrast to large cavities connected by narrow channels; we found that this generalizes to other classes of materials (Figure S23). We further looked at the *dimensionality* of the channels and found that the probability that a material is within the top 10% most selective materials given that it consists of one-dimensional channels is higher than for two- and three-dimensional channels for all classes of materials except for COFs (Section S16 for details). However, 1D channels are not required to form a highly selective binding site.

Our spherical and cylindrical shell models suggested that spherical cages would have higher selectivities than tube-like channels (Figure 7(b)). Inherent in this comparison is that the surface density of atoms η is equivalent in the two shell shapes; however, a spherical shell with a high surface density in all radial directions is not pragmatic, as an opening must exist for the adsorbate to enter the pocket. Also, there is an upper limit to the surface density η that can be obtained in real materials. Due to material design constraints, it may be that this surface density is easier to obtain with 1D tube-like channels.

Binding site analysis

To gain chemical intuition, we detected and analyzed the strongest Xe binding sites in the most selective materials by finding the minimum potential energy position of a xenon atom. Here, we analyze the molecular structure of the most selective binding sites.

First, we create a working definition of which atoms in the structure form the binding site. We expand a sphere centered at the binding site until 90% of the computed potential energy is accounted for (recall we employ a Lennard-Jones cutoff radius of 12.5 Å to define the binding energy); all atoms falling in this sphere are defined to create the binding site. For a selection of materials with the highest selectivities, we display a Xe atom (teal) centered at the binding site with the contributing structure atoms in Figure 10. The neighboring bar plot shows the cumulative contribution of each atom type to the binding energy as we expand the sphere. The vertical dashed line is the sphere radius such that 90% of the binding energy is accounted for, and it corresponds to the binding site visualization in the left panel.

Figure 10 shows that binding sites can be constructed from a striking diversity of geometries. For example, the binding site in one of the most selective materials, JAVTAC, is constructed inside a 1D channel (a) by forming a cylinder-like *tube* to host Xe atoms. The two structures in (b) form binding sites in a 1D channel using discrete chemical fragments, namely aromatic rings. Both structures suggest that aromatic rings provide a favorable arrangement of carbon atoms to achieve high Xe binding energies; note that C atoms account for most of the binding energy. While in (a) and (b), the structures form a *tube* of optimal diameter to accommodate Xe, the 1D channel in the MOF in (c) exhibits a *pocket* inside the channel with a favorable Xe binding energy. The hPPN in (d) also forms 1D channels, but it creates the binding site in a slightly different manner. With its adamantane core and benzonitrile monomer, it forms a *ring* that provides strong interactions with Xe. The hPPN in (e) exhibits 3D channels. In this hPPN, the binding site is surrounded by a *cage* of structure atoms in all directions.

In Section S17, we show the binding sites of the five most selective materials within each class, detailed visualizations of how the atoms contribute to the binding site, and atom counts. The most selective binding sites range from 43 to 193 atoms. In some cases, a small number of atoms are at the optimal interaction distance from the binding site; in other cases, a larger number of atoms are offset from the optimal interaction distance but achieve the a similarly favorable binding energy. We observed no clear correlation between the number of atoms forming the binding site and material family.

In summary, binding sites are complex entities, displaying a diversity of often non-discrete chemical fragments and arrangements in space. Based on this analysis of the top materials, we believe it will be difficult to rationally design crystalline materials with such sites embedded in their structures. This underscores the importance of using computational screenings to discover the top materials for Xe/Kr separations.

Conclusions

We computationally screened a set of 670,000 predicted crystal structures of zeolites, MOFs, ZIFs, PPNs, and COFs as well as already-synthesized zeolites and MOFs for promising materials for Xe/Kr separations at room temperature. The brute-force method of screening a database of materials by performing molecular simulations of Xe/Kr adsorption in each material is prohibitively expensive. To overcome this computational barrier to screening, we used a hybrid molecular simulation- machine learning approach. We characterized each of the structures by more easily-computed geometric properties, namely the largest included and free spheres, void fraction, surface area, and crystal density. We invented a new structural descriptor, the Voronoi energy, that is the average energy of a xenon atom at the Voronoi nodes of the accessible pore space. The Voronoi energy was found to be the most predictive structural descriptor. Using these structural descriptors

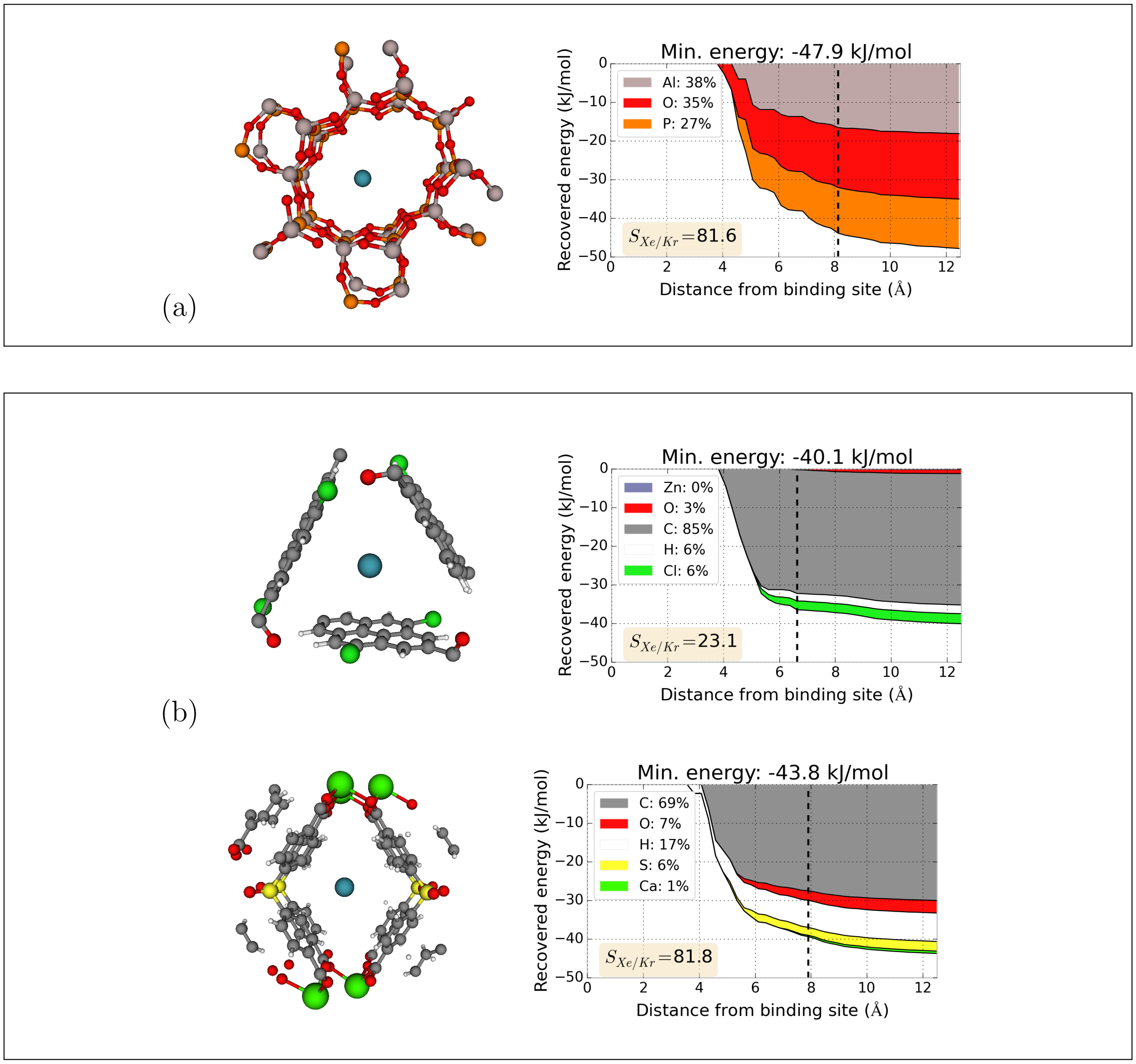

Min. energy: -47.9 kJ/mol
Al: 38%
O: 35%
P: 27%
$S_{Xe/Kr}$=81.6
Recovered energy (kJ/mol)
0
−10
−20
−30
−40
−50
0
2
4
6
8
10
12
Distance from binding site (Å)
(a)
Min. energy: -40.1 kJ/mol
Zn: 0%
O: 3%
C: 85%
H: 6%
Cl: 6%
$S_{Xe/Kr}$=23.1
Recovered energy (kJ/mol)
Distance from binding site (Å)
(b)
Min. energy: -43.8 kJ/mol
C: 69%
O: 7%
H: 17%
S: 6%
Ca: 1%
$S_{Xe/Kr}$=81.8
Recovered energy (kJ/mol)
Distance from binding site (Å)

Figure 10

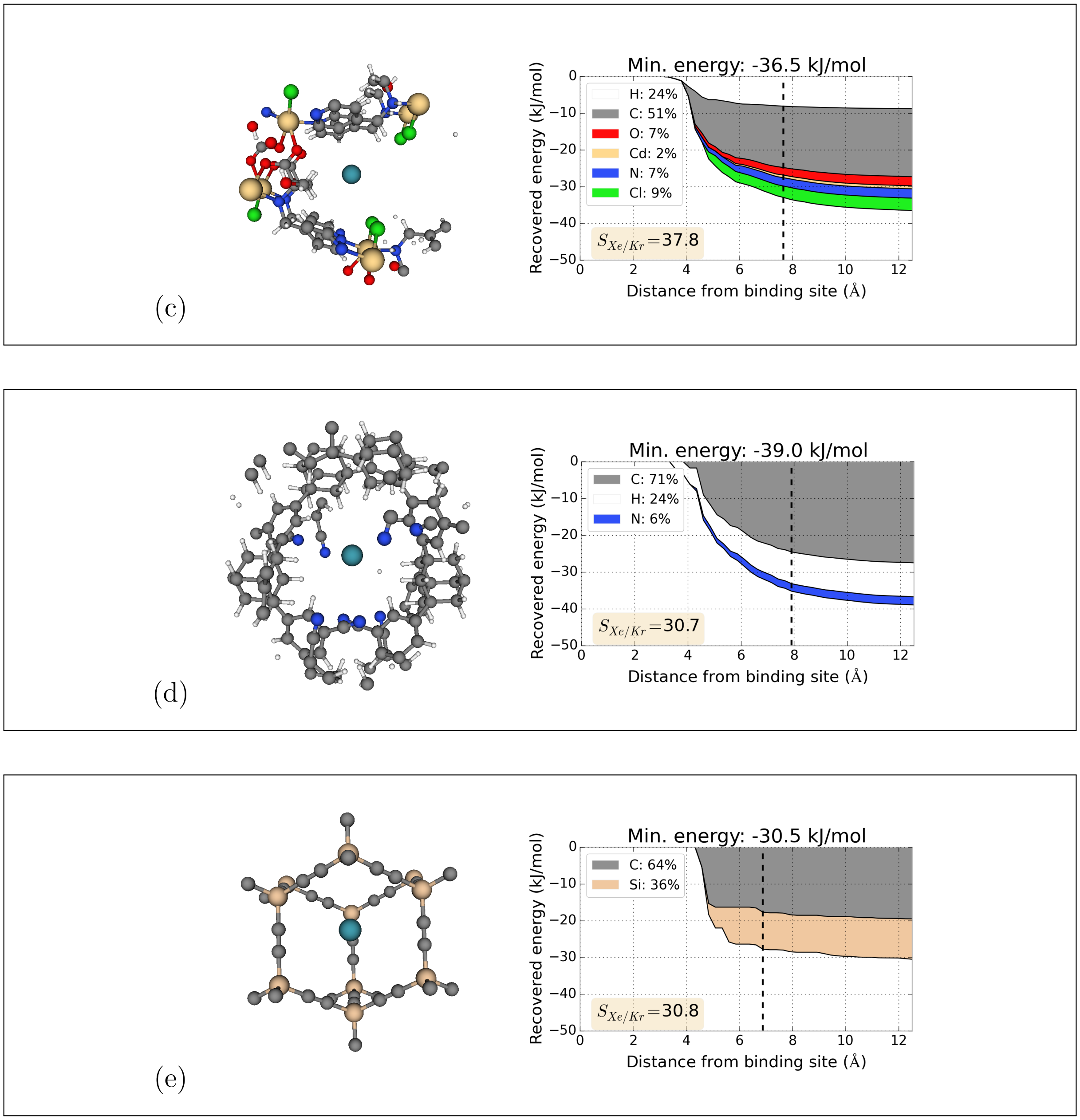


Figure 10: Anatomy of binding sites in a few of the most selective materials. (Left column) Xe atom (teal) at the minimum energy position with surrounding atoms contributing to the binding site. (Right column) The binding energy recovered as we include atoms at an increasing distance from the Xe. Colors show the contribution of each atom type to the binding energy; legend shows percentage of binding energy attributed to each atom type. Binding sites can be created by a tube-like arrangement of atoms along 1D channels (a,b). In (b), discrete chemical fragments – aromatic rings – form the binding site. Binding sites can also be *pocket*-like (c), *ring*-like (d), or *shell*-like (e). Structures: CoRE MOF JAVTAC,[105] hMOF_15308_i_0_j_17_k_17_m_2_cat_2, CoRE MOF KAXQIL,[110] CoRE MOF HEBKEG, hPPN_adamantane_2045_2-net_001, hPPN_Si_4080_1-net_001.

as a feature vector to serve as a fingerprint of the material in a high dimensional space, we trained a random forest of decision tree regressors to predict selectivity for xenon over krypton by showing the forest a mathematically diverse set of 15,000 training examples. For the remaining 655,000 materials, we predicted the selectivity by running their feature vectors through the trained random forest. We then used molecular simulations to refine the prediction if the random forest predicted the material to be promising. In this manner, we screened the Nanoporous Materials Genome by performing molecular simulations in only 20,000 of the 670,000 structures. Our screening strategy exploits the predictive power of random forests and is a promising high-throughput screening paradigm for expediting the discovery of new materials in the face of rapidly expanding materials datasets. Our screening paradigm can rapidly accelerate high-throughput screenings of the Nanoporous Materials Genome for other applications of porous materials, such as gas storage, gas sensing, drug delivery, and catalysis.

Many materials in our database are predicted to have better Xe/Kr separation performance than CC3,[104] a leading material for Xe/Kr separations.[44] Our models predict that the two most selective materials in the Nanoporous Materials Genome are JAVTAC, an aluminophosphate zeolite analogue,[105] and KAXQIL, a calcium coordination network.[110] Both materials have been synthesized but not yet tested for Xe/Kr separations. We hope that our open database of simulated Xe uptake and Xe/Kr selectivities (`http://nanoporousmaterials.org/xekrseparations/`) will inspire the synthesis and characterization of a new material for Xe/Kr separations.

Our analysis of the spherical and cylindrical shell models rationalized the strong link between the selectivity and the pore size. Still, by comparing the structural descriptors of good-performing to poor-performing materials, we found that there is not a simple recipe of geometric descriptors that will guarantee a material to be good for Xe/Kr separations.

This motivates using machine learning algorithms to learn the relationship between selectivity and features in a high dimensional space.

By analyzing the top materials, we revealed chemical insights about their highly selective binding sites. They can be constructed from a diverse array of often non-discrete chemical fragments, arranged into tubes, cages, pockets, or rings. The complexity of these binding sites makes rational design difficult and underscores the importance of high-throughput screening for the discovery of novel and/or optimal materials for a given application.

Acknowledgements

C.M.S. is supported by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research (SCGSR) program. The SCGSR program is administered by the Oak Ridge Institute for Science and Education for the DOE under contract number DE-AC05-06OR23100. R.M. is supported by the UC Berkeley Chancellor's Fellowship. S.K.S. is supported by the Research Council of Norway through a post-doc fellowship, grant no. 230534. B.S. was supported by the Center for Gas Separations Relevant to Clean Energy Technologies, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award number DE-SC0001015. M.H. acknowledge U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences under Award DE-FG02-12ER16362. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Thank you Linjiang Chen for performing the sim-

ulation of Xe/Kr adsorption in the porous cage material CC3. Thank you Stephanie Teich-McGoldrick for kindly providing a simulation-ready IRMOF-2 crystal structure.

Supporting Information Available

Supporting information is available for more plots and details alluded to in the main text. This information is available free of charge via the Internet at http://pubs.acs.org/. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Morris, R.; Wheatley, P. S. *Angew. Chem., Int. Ed.* **2008**, *47*, 4966–4981.

(2) Snurr, R. Q.; Hupp, J. T.; Nguyen, S. T. *AIChE J.* **2004**, *50*, 1090–1095.

(3) Kreno, L. E.; Leong, K.; Farha, O. K.; Allendorf, M.; Van Duyne, R. P.; Hupp, J. T. *Chem. Rev.* **2011**, *112*, 1105–1125.

(4) Lee, J.; Farha, O. K.; Roberts, J.; Scheidt, K. A.; Nguyen, S. T.; Hupp, J. T. *Chem. Soc. Rev.* **2009**, *38*, 1450–1459.

(5) Horcajada, P.; Serre, C.; Vallet-Regí, M.; Sebban, M.; Taulelle, F.; Férey, G. *Angew. Chem., Int. Ed.* **2006**, *45*, 5974–5978.

(6) Zhou, H.-C.; Long, J. R.; Yaghi, O. M. *Chem. Rev.* **2012**, *112*, 673–674.

(7) Côté, A. P.; Benin, A. I.; Ockwig, N. W.; O'Keeffe, M.; Matzger, A. J.; Yaghi, O. M. *Science* **2005**, *310*, 1166–1170.

(8) Park, K. S.; Ni, Z.; Côté, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O'Keeffe, M.; Yaghi, O. M. *Proc. Natl. Acad. Sci.* **2006**, *103*, 10186–10191.

(9) Lu, W.; Yuan, D.; Zhao, D.; Schilling, C. I.; Plietzsch, O.; Muller, T.; Bräse, S.; Guenther, J.; Blümel, J.; Krishna, R.; Li, Z.; Zhou, H.-C. *Chem. Mater.* **2010**, *22*, 5964–5972.

(10) Materials Genome Initiative for Global Competitiveness. 2011; `http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf`.

(11) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. *APL Mater.* **2013**, *1*, 011002.

(12) Dobson, C. M. *Nature* **2004**, *432*, 824–828.

(13) Nanoporous Materials Genome Center. `http://www.chem.umn.edu/nmgc/`.

(14) Wilmer, C. E.; Leaf, M.; Lee, C.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q. *Nat. Chem.* **2011**, *4*, 83–89.

(15) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. *J. Phys. Chem. C* **2009**, *113*, 21353–21360.

(16) Pophale, R.; Cheeseman, P. A.; Deem, M. W. *Phys. Chem. Chem. Phys.* **2011**, *13*, 12407–12412.

(17) Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M. *J. Am. Chem. Soc.* **2014**, *136*, 5006–5022.

(18) Lin, L.-C.; Berger, A. H.; Martin, R. L.; Kim, J.; Swisher, J. A.; Jariwala, K.; Rycroft, C. H.; Bhown, A. S.; Deem, M. W.; Haranczyk, M.; Smit, B. *Nat. Mater.* **2012**, *11*, 633–641.

(19) Martin, R.; Simon, C.; Medasani, B.; Britt, D.; Smit, B.; Haranczyk, M. *J. Phys. Chem. C* **2014**, *19*, 186–195.

(20) *International Zeolite Association* `http://www.iza-structure.org/databases/`.

(21) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. *Chem. Mater.* **2014**, *26*, 6185–6192.

(22) Simon, C.; Kim, J.; Gomez-Gualdron, D.; Camp, J.; Chung, Y. G.; Martin, R. L.; Mercado, R.; Deem, M. W.; Gunter, D.; Haranczyk, M.; Sholl, D.; Snurr, R. Q.; Smit, B. *Energy Environ. Sci.* **2015**, http://dx.doi.org/10.1039/C4EE03515A.

(23) Wilmer, C. E.; Farha, O. K.; Bae, Y.-S.; Hupp, J. T.; Snurr, R. Q. *Energy Environ. Sci.* **2012**, *5*, 9849–9856.

(24) Colón, Y. J.; Fairen-Jimenez, D.; Wilmer, C. E.; Snurr, R. Q. *J. Phys. Chem. C* **2014**, *118*, 5383–5389.

(25) Kim, J.; Lin, L.-C.; Martin, R. L.; Swisher, J. A.; Haranczyk, M.; Smit, B. *Langmuir* **2012**, *28*, 11914–11919.

(26) Bai, P.; Jeon, M. Y.; Ren, L.; Knight, C.; Deem, M. W.; Tsapatsis, M.; Siepmann, J. I. *Nat. Commun.* **2015**, *6*.

(27) Matito-Martos, I.; Martin-Calvo, A.; Gutierrez-Sevillano, J. J.; Haranczyk, M.; Doblare, M.; Parra, J. B.; Ania, C. O.; Calero, S. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19884–19893.

(28) Air Liquide. `http://www.airliquide.com/en/company/our-businesses-our-products/air-gases/noble-gases-krypton-neon-xenon/noble-gases-applications.html`.

(29) Cullen, S. C.; Gross, E. G. *Science* **1951**, *113*, 580–582.

(30) Franks, N. P.; Dickinson, R.; de Sousa, S. L. M.; Hall, A. C.; Lieb, W. R. *Nature* **1998**, *396*, 324–324.

(31) Sanders, R. D.; Ma, D.; Maze, M. *Brit. Med. Bull.* **2005**, *71*, 115–135.

(32) Albert, M.; Cates, G.; Driehuys, B.; Happer, W.; Saam, B.; Springer, C.; Wishnia, A. *Nature* **1994**, *370*, 199–201.

(33) Beattie, J.; Matossian, J.; Poeschel, R.; Rogers, W.; Martinelli, R. *J. Propul. Power* **1989**, *5*, 438–444.

(34) Yeralan, S.; Doughty, D.; Blondia, R.; Hamburger, R. *Proc. SPIE* **2005**, *5740*, 27–35.

(35) Bridges, W. B.; Chester, A. N. *Appl. Optics* **1965**, *4*, 573–580.

(36) Hoff, P. W.; Swingle, J. C.; Rhodes, C. K. *Appl. Phys. Lett.* **1973**, *23*, 245–246.

(37) Manz, H. *Renew. Energ.* **2008**, *33*, 119 – 128.

(38) Weir, G.; Muneer, T. *Energ. Convers. Manage.* **1998**, *39*, 243 – 256.

(39) Lipsky, S.; Shahin, M. *Nature* **1963**, *200*, 566 – 567.

(40) Hwang, S.-C.; Weltmer, W. R. *Kirk-Othmer Encyclopedia of Chemical Technology*; John Wiley & Sons, Inc., 2000.

(41) Air Liquide. `http://www.airliquide.com/en/company/our-businesses-our-products/air-gases/noble-gases-krypton-neon-xenon.html`.

(42) Kerry, F. G. *Industrial Gas Handbook: Gas Separation and Purification*; CRC Press, 2010.

(43) Sikora, B. J.; Wilmer, C. E.; Greenfield, M. L.; Snurr, R. Q. *Chem. Sci.* **2012**, *3*, 2217–2223.

(44) Chen, L. et al. *Nat. Mater.* **2014**, *13*, 954–960.

(45) Smit, B.; Reimer, J. R.; Oldenburg, C. M.; Bourg, I. C. *Introduction to Carbon Capture and Sequestration*; Imperial College Press, London, 2014.

(46) Sircar, S. *Ind. Eng. Chem. Res.* **2002**, *41*, 1389–1392.

(47) Banerjee, D.; Cairns, A. J.; Liu, J.; Motkuri, R. K.; Nune, S. K.; Fernandez, C. A.; Krishna, R.; Strachan, D. M.; Thallapally, P. K. *Acc. Chem. Res.* **2014**, DOI: 10.1021/ar5003126.

(48) Liu, J.; Thallapally, P. K.; Strachan, D. *Langmuir* **2012**, *28*, 11584–11589.

(49) Izumi, J. *Handbook of Zeolite Science and Technology*; Marcel Dekker; New York, Basel, 2003.

(50) Bazan, R.; Bastos-Neto, M.; Moeller, A.; Dreisbach, F.; Staudt, R. *Adsorption* **2011**, *17*, 371–383.

(51) Munakata, K.; Fukumatsu, T.; Yamatsuki, S.; Tanaka, K.; Nishikawa, M. *J. Nucl. Sci. Technol.* **1999**, *36*, 818–829.

(52) Foroutan, M.; Nasrabadi, A. T. *Chem. Phys. Lett.* **2010**, *497*, 213–217.

(53) Meek, S. T.; Teich-McGoldrick, S. L.; Perry, J. J.; Greathouse, J. A.; Allendorf, M. D. *J Phys. Chem. C* **2012**, *116*, 19765–19772.

(54) Parkes, M. V.; Staiger, C. L.; Perry IV, J. J.; Allendorf, M. D.; Greathouse, J. A. *Phys. Chem. Chem. Phys.* **2013**, *15*, 9093–9106.

(55) Wang, H.; Yao, K.; Zhang, Z.; Jagiello, J.; Gong, Q.; Han, Y.; Li, J. *Chem. Sci.* **2014**, *5*, 620–624.

(56) Thallapally, P. K.; Grate, J. W.; Motkuri, R. K. *Chem. Commun.* **2012**, *48*, 347–349.

(57) Fernandez, C. A.; Liu, J.; Thallapally, P. K.; Strachan, D. M. *J. Am. Chem. Soc.* **2012**, *134*, 9046–9049.

(58) Liu, J.; Strachan, D. M.; Thallapally, P. K. *Chem. Commun.* **2014**, *50*, 466–468.

(59) Lawler, K. V.; Hulvey, Z.; Forster, P. M. *Chem. Commun.* **2013**, *49*, 10959–10961.

(60) Mueller, U.; Schubert, M.; Teich, F.; Puetter, H.; Schierle-Arndt, K.; Pastre, J. *J. of Mater. Chem.* **2006**, *16*, 626–636.

(61) Bae, Y.-S.; Hauser, B. G.; Colón, Y. J.; Hupp, J. T.; Farha, O. K.; Snurr, R. Q. *Microporous Mesoporous Mater.* **2013**, *169*, 176–179.

(62) Magdysyuk, O.; Adams, F.; Liermann, H.-P.; Spanopoulos, I.; Trikalitis, P.; Hirscher, M.; Morris, R.; Duncan, M. J.; McCormick, L.; Dinnebier, R. E. *Phys. Chem. Chem. Phys.* **2014**, *16*, 23908–23914.

(63) Ryan, P.; Farha, O. K.; Broadbelt, L. J.; Snurr, R. Q. *AIChE J.* **2011**, *57*, 1759–1766.

(64) Van Heest, T.; Teich-McGoldrick, S. L.; Greathouse, J. A.; Allendorf, M. D.; Sholl, D. S. *J. Phys. Chem. C* **2012**, *116*, 13183–13195.

(65) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. *Chem. Rev.* **2012**, *112*, 2889–2919.

(66) Breiman, L. *Mach. Learn.* **2001**, *45*, 5–32.

(67) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273–297.

(68) Hastie, T.; Tibshirani, R.; Friedman, J.; Hastie, T.; Friedman, J.; Tibshirani, R. *The elements of statistical learning*; Springer, 2009; Vol. 2.

(69) Fernandez, M.; Woo, T. K.; Wilmer, C. E.; Snurr, R. Q. *J. Phys. Chem. C* **2013**, *117*, 7681–7689.

(70) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. *J. Phys. Chem. Lett.* **2014**, *5*, 3056–3060.

(71) Dietterich, T. G. *Multiple Classifier Systems*; Lecture Notes in Computer Science; Springer Berlin Heidelberg, 2000; Vol. 1857; pp 1–15.

(72) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT Press, 2012.

(73) Surowiecki, J. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business*; Doubleday; Anchor, 2004.

(74) Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. Proc. Int. Conf. Machine Learn., 23rd, Pittsburgh, PA. 2006; pp 161–168.

(75) Shotton, J.; Sharp, T.; Kipman, A.; Fitzgibbon, A.; Finocchio, M.; Blake, A.; Cook, M.; Moore, R. *Commun. ACM* **2013**, *56*, 116–124.

(76) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137–148.

(77) Boato, G.; Casanova, G. *Physica* **1961**, *27*, 571 – 589.

(78) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard Iii, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(79) Talu, O.; Myers, A. L. *Colloids Surf., A* **2001**, *187-188*, 83–93.

(80) Coudert, F.-X. *Chem. Mat.* **2015**, *27*, 190–1916.

(81) Sarkisov, L.; Martin, R. L.; Haranczyk, M.; Smit, B. *J. Am. Chem. Soc.* **2014**, *136*, 2228–2231.

(82) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: from Algorithms to Applications*, 2nd ed.; Academic Press, San Diego, 2002; Vol. 1.

(83) Kim, J.; Smit, B. *J. Chem. Theory Comput.* **2012**, *8*, 2336–2343.

(84) Golden, T.; Sircar, S. *J. Colloid Interface Sci.* **1994**, *162*, 182–188.

(85) Perry IV, J. J.; Teich-McGoldrick, S. L.; Meek, S. T.; Greathouse, J. A.; Haranczyk, M.; Allendorf, M. D. *J Phys. Chem. C* **2014**,

(86) McDaniel, J. G.; Li, S.; Tylianakis, E.; Snurr, R. Q.; Schmidt, J. R. *J. Phys. Chem. C* **2015**, null.

(87) Martin, R. L.; Prabhat, M.; Donofrio, D. D.; Sethian, J. A.; Haranczyk, M. *Int. J. High Perform. Comput.* **2012**, *26*, 347–357.

(88) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. *Microporous Mesoporous Mater.* **2012**, *149*, 134–141.

(89) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(90) Rowland, R. S.; Taylor, R. *J. Phys. Chem.* **1996**, *100*, 7384–7391.

(91) Rycroft, C. H. *Chaos* **2009**, *19*, 041111.

(92) Li, H.; Laine, A.; O'Keeffe, M.; Yaghi, O. M. *Science* **1999**, *283*, 1145–1147.

(93) Martin, R.; Willems, T.; Lin, L.; Kim, J.; Swisher, J.; Smit, B.; Haranczyk, M. *ChemPhysChem* **2012**, *13*, 3595–3597.

(94) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and regression trees*; CRC Press, 1984.

(95) Pedregosa, F. et al. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(96) Criminisi, A.; Shotton, J.; Konukoglu, E. *Foundations and Trends® in Computer Graphics and Vision* **2012**, *7*, 81–227.

(97) Breiman, L. *Mach. Learn.* **1996**, *24*, 123–140.

(98) Dietterich, T. G. *Mach. Learn.* **2000**, *40*, 139–157.

(99) Ho, T. K. *IEEE Trans. Pattern Anal. Mach. Intell* **1998**, *20*, 832–844.

(100) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. *BMC Bioinform.* **2008**, *9*, 307.

(101) Cox, B.; Thamwattana, N.; Hill, J. *Proc. R. Soc. London, Ser. A* **2007**, *463*, 477–494.

(102) Ripmeester, J.; Ratcliffe, C. *J. Phys. Chem.* **1990**, *94*, 7652–7656.

(103) Thamwattana, N.; Baowan, D.; Cox, B. J. *RSC Advances* **2013**, *3*, 23482–23488.

(104) Tozawa, T. et al. **2009**, *8*, 973–978.

(105) Cooper, E. R.; Andrews, C. D.; Wheatley, P. S.; Webb, P. B.; Wormald, P.; Morris, R. E. *Nature* **2004**, *430*, 1012–1016.

(106) Haranczyk, M.; Lin, L.-C.; Lee, K.; Martin, R. L.; Neaton, J. B.; Smit, B. *Phys. Chem. Chem. Phys.* **2013**, *15*, 20937–20942.

(107) Wilson, S. T.; Lok, B. M.; Messina, C. A.; Cannan, T. R.; Flanigen, E. M. *J. Am. Chem. Soc.* **1982**, *104*, 1146–1147.

(108) Song, X.; Li, J.; Guo, Y.; Pan, Q.; Gan, L.; Yu, J.; Xu, R. *Inorg. Chem.* **2009**, *48*, 198–203.

(109) Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley & Sons, New York, Chichester, 1992.

(110) Banerjee, D.; Zhang, Z.; Plonka, A. M.; Li, J.; Parise, J. B. *Crystal Growth & Design* **2012**, *12*, 2162–2165.

Graphical TOC Entry