JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

Addressing Challenges of Identifying Geometrically Diverse Sets of Crystalline Porous Materials

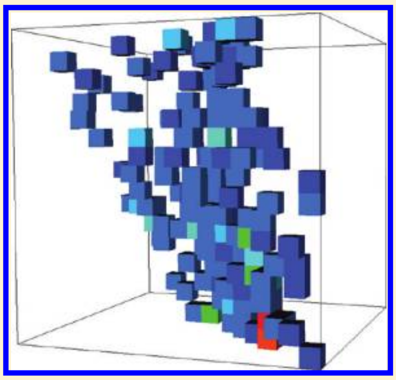
Richard Luis Martin,[†] Berend Smit,[‡,§] and Maciej Haranczyk*,[†]

[†]Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, California 94720-8139, United States

[‡]Department of Chemistry, University of California, Berkeley, California 94720-1462, United States

[§]Department of Chemical Engineering, University of California, Berkeley, California 94720-1462, United States

Supporting Information



ABSTRACT: Crystalline porous materials have a variety of uses, such as for catalysis and separations. Identifying suitable materials for a given application can, in principle, be done by screening material databases. Such a screening requires automated high-throughput analysis tools that calculate topological and geometrical parameters describing pores. These descriptors can be used to compare, select, group, and classify materials. Here, we present a descriptor that captures shape and geometry characteristics of pores. Together with proposed similarity measures, it can be used to perform diversity selection on a set of porous materials. Our representations are histogram encodings of the probe-accessible fragment of the Voronoi network representing the void space of a material. We discuss and demonstrate the application of our approach on the International Zeolite Association (IZA) database of zeolite frameworks and the Deem database of hypothetical zeolites, as well as zeolitic imidazolate frameworks constructed from IZA zeolite structures. The diverse structures retrieved by our method are complementary to those expected by emphasizing diversity in existing one-dimensional descriptors, e.g., surface area, and similar to those obtainable by a (subjective) manual selection based on materials' visual representations. Our technique allows for reduction of large sets of structures and thus enables the material researcher to focus efforts on maximally dissimilar structures.

INTRODUCTION

Porous materials contain complex networks of void channels and cages that are exploited in many industrial applications. The zeolite class of these materials is the most well-known, as they have found wide use in industry since the late 1950s, with common applications as chemical catalysts and membranes for separations and water softeners;[1−4] their value is estimated at $350 billion per year.[5] There is increasing interest in utilizing zeolites as membranes or adsorbents for CO_2 capture applications.[3] In addition to zeolites, metal organic frameworks (MOFs)[6,7] and their subfamily of zeolitic imidazolate frameworks (ZIFs)[8] have recently generated interest for their potential use in gas separation or storage.[9−11] A key requirement for the success of any nanoporous material is that the chemical composition and pore geometry and topology must be optimal under the given conditions for a particular application. However, finding the optimal material is an arduous task, since the number of possible pore topologies is extremely large. There are approximately 190 unique zeolite frameworks known to exist today in more than 1400 zeolite crystals of various chemical compositions and different geometrical parameters (see ref 12). However, these experimentally known zeolites constitute only a very small fraction of more than 2.7 million structures that are feasible on theoretical grounds.[13,14] Of these, between 314 000 and 585 000 structures are predicted to be thermodynamically accessible as aluminosilicates, which gives an even larger number of possible materials via elemental substitution and different cation exchanges.[15,16] Databases of similar or greater magnitude can be developed for other nanoporous materials such as MOFs or ZIFs. As a result, new automated computational and cheminformatic techniques need to be developed to characterize, categorize, and screen such large databases.[17]

Recently, automated approaches capable of performing analysis of large sets of porous materials have started to emerge. For example, Blatov and co-workers have pursued the concept of natural tiling of periodic networks to find primitive building blocks in zeolites.[18] The group of Blaisten-Barojas has developed zeolite framework classifiers using a machine learning approach.[19] Düren et al. have provided a tool to calculate the surface area of a porous material,[20] while Foster et al. and Haldoupis et al. have presented methods to calculate two parameters frequently used to describe pore geometry in crystalline porous materials,[17,21] namely, the diameter of the largest included (d_i) and the largest

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: August 15, 2011
Published: November 18, 2011

free (d_f) spheres.[22] The largest included sphere points to the location of the largest cavity in a porous material and measures its size. In contrast, the largest free sphere corresponds to the largest spherical probe that can diffuse through a structure and measures a minimum restricting aperture on a diffusion path. Using their method to calculate d_f, Haldoupis et al. analyzed a hypothetical zeolite database containing more than 250 000 structures[23] as well as hundreds of MOFs and directed a few thousand of them for further characterization using molecular simulation methods.

Although Haldoupis et al. have pushed the current limits in terms of the number of investigated porous materials, their method's reliance on a single simplistic structural descriptor, d_f, demonstrates the narrow range of descriptors presently available in state-of-the-art structural representation. There is a need therefore to develop additional, novel structural descriptors, as well as further expand upon the range of available tools and approaches for structure analysis, selection, comparison, and investigation of the geometrical parameters describing pores. Recently, the Floudas group has begun to address this issue: they developed an automatic approach to segment and analyze the void space of zeolites[24] as well as proposed a screening approach for materials with shape selectivity.[25] Our group has also begun to address this issue[26−28] and porous materials-specific visualization needs.[29,30] We presented algorithms and software tools for high-throughput geometry-based analysis of crystalline porous materials, in particular, efficient algorithms to calculate d_i and d_f for a given structure as well as the dimensionality of its channel systems.[26] Moreover, we provided algorithms to determine the accessibility of sections of the void space to a particular probe, as well as a Monte Carlo procedure for integration of accessible surface area (ASA) and accessible volume (AV) that can use the resulting information. Our tools are based on the Voronoi decomposition, which for a given arrangement of atoms in a periodic domain provides a graph representation of the void space. When performing a Voronoi decomposition, the space surrounding n points is divided into n polyhedral cells such that each of their faces lies on the plane equidistant from the two points sharing the face. Edges of such cells overlap with lines equidistant to neighboring points (three points in a general asymmetric case), whereas vertices of cells, the Voronoi nodes, are equidistant from neighboring points (four points in a general asymmetric case). The Voronoi network, built of such nodes and edges, maps the void space surrounding the points. Analysis of such a network is fairly straightforward and can provide detailed information about void space geometry and topology. The Voronoi decomposition has already been used in the analysis of crystalline materials[31] and their voids[32] as well as membranes[33] and has been suggested as a tool to investigate ion transport pathways in crystals.[34]

The vast majority of currently available descriptors, such as d_f, d_i, ASA, and AV, are one-dimensional descriptors. As such they have a limited application in diversity selection of materials with various shapes and geometries of pores. Such selection is in demand for at least two reasons. First, large databases containing millions of hypothetical material structures are becoming available, and the computational cost of their characterization using molecular simulation techniques can be prohibitively high. Efficient sampling ensures that valuable resources are spent on statistically relevant, nonidentical structures. Second, with state-of-the-art molecular simulation studies characterizing hundreds of thousands of materials, the quantity of "top candidates" can still be too large to allow visualization or more detailed structural analysis of each material. Instead, the researcher can examine a much smaller quantity of high-performance materials, and within this maximally dissimilar set, discover specific features that they have in common.

In the present contribution, we introduce a new geometry-based descriptor that aims to capture shape and size characteristics of the accessible sections of the void space. Our descriptor, the Voronoi hologram, is a three-dimensional vector holding a histogram representation of the accessible section of a Voronoi network encoding of the void space of a material. This representation is efficiently combined with a modified Tanimoto similarity coefficient and dissimilarity-based selection algorithms to perform diversity selection of porous materials. We present an application of our approach on the International Zeolite Association (IZA) database of zeolite topologies, as well as Deem's database of hypothetical zeolites[16] and a set of computationally derived ZIF structures.

■ METHODS

Overview. Our approach to diversity selection consists of the following steps: (1) Perform the Voronoi decomposition for all structures in the data set to obtain a graph representation of the void space of materials. (2) Obtain Voronoi hologram representations of the graphs of 1. (3) Initiate the diversity selection procedure by selecting the first structure. (4) Iteratively identify the most diverse structure using a similarity measure, and add the structure to the sample, continuing until the similarity threshold is reached.

In the following sections, we discuss the details of components used to execute steps 1−4.

1. Voronoi Network Representation of the Accessible Void Space. Our implementation of the Voronoi decomposition in analysis of the void space of crystalline porous materials has been described in detail in ref 26.

The Voronoi decomposition for a particular porous material m yields a periodic Voronoi network, w, which is a function of the spatial arrangement of atoms. w consists of nodes and edges mapping the void space surrounding the atoms (Figure 1), where each node and edge is labeled with its distance to the nearest atoms. This distance corresponds to the radii of the largest spherical probe that can, respectively, be placed at the node or travel along the edge, without colliding with any atom. For a particular probe radius (for example, 1.625 Å for CH_4 probe), a graph propagation algorithm—a variation of the Dijkstra shortest path algorithm[35]—is then used to identify the probe-accessible regions of the Voronoi network, v. v is a periodic subgraph representing the guest-molecule-accessible region of the void space. Voronoi networks obtained for different materials can be compared to produce a measure of shape similarity between the void space networks of materials. It has to be noted that our implementation of the Voronoi decomposition can handle atoms of different radii.[26] The approach discussed here can therefore be applied to diverse materials such as MOFs and ZIFs. For the purpose of this article, we have used the same set of radii as in previous studies of zeolites[21] and MOFs.[17]

2. Voronoi Hologram Representations. Given a probe-molecule-accessible Voronoi network v, we construct histograms which encode the frequency of occurrence of edges within v, classified by three properties: their length, l, and the radii of the two nodes they connect, r_a and r_b, where $r_a \geq r_b$. The interpretation of radius is simply the distance from the Voronoi node to the surface of the

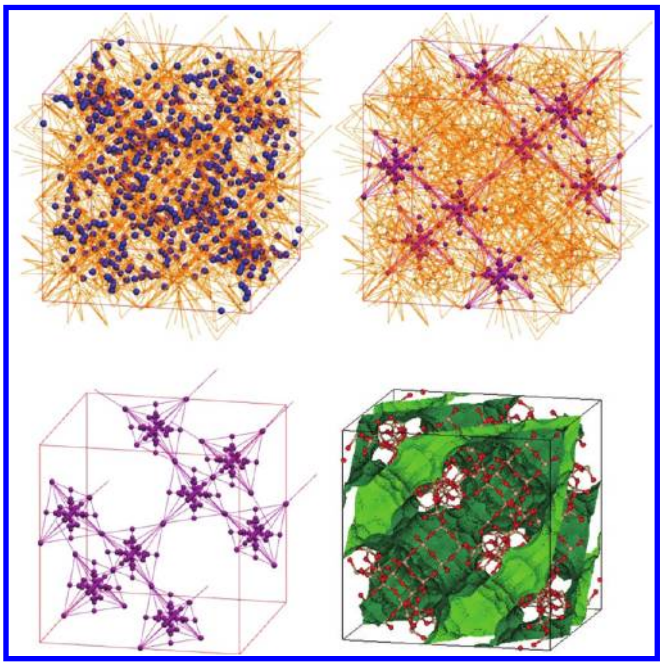


Figure 1. Top left: an example material m—the zeolite FAU (atoms in blue)—and its 3D Voronoi network w (orange). Top right: within w, the CH_4-accessible Voronoi subnetwork v is highlighted (purple). Bottom left: only v is shown, to illustrate the pore topology encoded as Voronoi nodes and edges. Bottom right: a visualization of the pore landscape corresponding to the CH_4-accessible network, with the Si and O atoms in the structure in tan and red, respectively. Atom and node radii are not shown.

Table 1. Edge Length and Node Radii Upper Bounds, in Ångstroms, to Three Decimal Places[a]

bin	edge length	node radii
1	0.060	1.725
2	0.139	1.827
3	0.211	1.932
4	0.308	2.029
5	0.407	2.105
6	0.518	2.193
7	0.617	2.310
8	0.733	2.423
9	0.860	2.520
10	1.018	2.650
11	1.196	2.783
12	1.437	2.975
13	1.757	3.173
14	2.231	3.400
15	3.081	3.835

[a] There is no upper bound for the 16th bin, since it includes all occurrences of lengths/radii above the bound of the 15th bin.

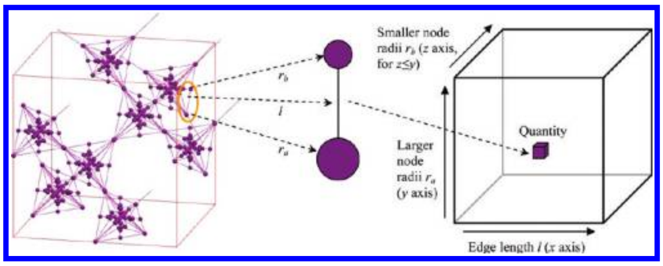

Smaller node radii r_b (z axis, for $z \leq y$)
r_b
l
r_a
Larger node radii r_a (y axis)
Quantity
Edge length l (x axis)

Figure 2. The creation of a Voronoi hologram, $h(v)$, for the zeolite FAU. (1) Probe-accessible channels in v are detected (left, for clarity of presentation, nodes are visualized with equal and small radii). (2) Each edge has a binned length and two node radii (center). (3) The quantity of occurrence of bin combination is encoded in a discrete 3D grid (right).

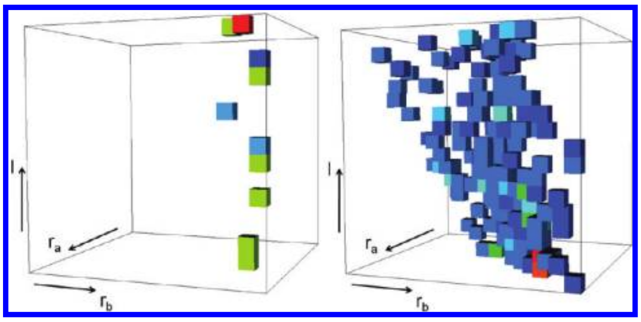


Figure 3. Left: the Voronoi hologram for zeolite FAU. Features are assigned a color on the basis of their frequency of occurrence, ranging in this case from 64 (dark blue) to 256 (red). Right: the Voronoi hologram for zeolite TUN, the densest in the IZA set. The frequency of feature occurrence in this case ranges from 4 (dark blue) to 104 (red).

nearest atom given by the Voronoi decomposition procedure. Because l, r_a, and r_b are continuous variables, we chose to construct a binning system for edge lengths and node radii. The bounds for each bin were determined by profiling a data set of combined CH_4-accessible IZA (148 structures) and Deem (200, randomly selected structures) databases. In this, we emphasize equally populated bins; as such, the bin upper bounds can be said to be tuned to this random selection of structures. We found that the use of 16 bins provides an approximate average length step size of 0.2 Å and a radii step size of 0.25 Å, and we selected this quantity. The upper bounds are provided in Table 1. The first bound for edge lengths is quite small, indicating the high quantity of very short edges present in these networks; note also that the smallest node radii which can occur in the accessible part of a network will be equal to the probe radius—in this case, CH_4 with radius of 1.625 Å. The result is that each edge in v belongs to exactly one of the 2176 distinct bins arranged in a three-dimensional, cubic grid (note that the region of the grid representing $r_a < r_b$ is unoccupied by the above definition). Hence, v is uniquely represented by $h(v)$, where $h(v)$ is a concise encoding of the multiplicity of edge types which occur in v. The construction of $h(v)$ is illustrated in Figure 2; two example holograms, those of the zeolites FAU and TUN, are provided in Figure 3.

$h(v)$ constitutes an abstraction, or simplification, of v. Each edge e in v is present in $h(v)$; however, all further information—about the interconnectivity of edges, their position in space, etc.—is lost. With this lower-complexity representation, we can more efficiently compare two structures A and B. We compute their probe-accessible Voronoi networks v_A and v_B and measure the similarity between $h(v_A)$ and $h(v_B)$. The methods for calculating this similarity are discussed in step 4.

3. Diversity Selection. Diversity selection is a technique which allows for efficient selection of dissimilar structures from a large set.

The approach identifies least-similar structures out of a set using a chosen structure representation and the corresponding similarity measure.

With very large data sets, a pairwise comparison of structures becomes computationally expensive, and so we perform a MaxMin[36] maximum-dissimilarity-based selection, which is considered to be a highly effective method for the selection of diverse and representative samples[37] and which does not require a precalculated pairwise similarity matrix. The MaxMin method proceeds as follows:

(1) Initialize the method by selecting some starting structure (seed).
(2) For each remaining structure, determine its similarity to every structure which has been selected, storing the maximum observed similarity, *s*.
(3) Add the structure which exhibits the smallest *s* to the set of selected structures.
(4) If the specified end criteria have not been met, go to step 2.

By this method, it follows that the values of *s* observed at each step comprise a nondecreasing series, and so the method can be terminated once some similarity threshold is reached, or alternatively once a specified quantity of structures have been selected. In our implementation, the first structure in the alphabetically ordered list of structures was arbitrarily selected as the seed.

4. Similarity Coefficients. There exist many similarity coefficients for the calculation of similarity between binary (fingerprint) or nonbinary (hologram) arrays *A* and *B*. In this work, only holograms are generated; however, they can be compared in a binary manner through the application of a binary similarity coefficient. A commonly applied binary similarity coefficient[38] is the Tanimoto coefficient, $\mathrm{Tan_{bin}}$:

$$\mathrm{Tan_{bin}} = \frac{c}{a + b - c} \quad (1)$$

where *a* and *b* are the number of active (i.e., set to 1 in binary arrays, or nonzero in continuous arrays) bits in arrays *A* and *B*, and *c* is the number of active bits in common. It has a range from 0 (maximal dissimilarity) to 1 (identity). The continuous version of the Tanimoto similarity coefficient, $\mathrm{Tan_{cont}}$ (for nonbinary data), is given by:

$$\mathrm{Tan_{cont}} = \frac{\sum_{i=1}^{N} x_{iA}x_{iB}}{\sum_{i=1}^{N} (x_{iA})^2 + \sum_{i=1}^{N} (x_{iB})^2 - \sum_{i=1}^{N} x_{iA}x_{iB}} \quad (2)$$

where x_{iA} and x_{iB} represent the *i*th elements in arrays *A* and *B*. The similarity to $\mathrm{Tan_{bin}}$ (see eq 1) is clear, and as long as *A* and *B* do not contain negative entries, the range remains 0−1 as above.

For the purposes of this discussion, it is also useful to define how the binary Tanimoto coefficient can measure similarity on the basis of the common absence of features, rather than their common presence as in $\mathrm{Tan_{bin}}$. We denote this by $\mathrm{TanAbsence_{bin}}$:

$$\mathrm{TanAbsence_{bin}} = \frac{n + c - a - b}{n - c} \quad (3)$$

where *n* is the length of the arrays. Note that unlike common presence as seen above, it is not possible to measure common absence in a continuous manner.

We are interested in the construction of a representative sample of a data set of structures through diversity-based selection. However, the Tanimoto coefficient has a known bias toward the retrieval of simplistic structures in diversity-based selection procedures.[39] This is because for any binary fingerprint there may exist more than one (specifically, 2^{n-a}) distinct, and maximally dissimilar, fingerprint. This effect can be mitigated by considering, in addition to the common presence of features, *f*, the common absence thereof.[39] The modified Tanimoto system with this functionality devised by Fligner et al.[39] is referred to here as the binary modified Tanimoto with weighting, $\mathrm{MTW_{bin}}$:

$$\mathrm{MTW_{bin}} = \frac{1}{3}(\mathrm{Tan_{bin}}(2 - p) + \mathrm{TanAbsence_{bin}}(1 + p)) \quad (4)$$

where *p* denotes the proportion of nonzero bits in the combined arrays, given by:

$$p = \frac{a + b}{2n} \quad (5)$$

We modify this coefficient to produce a binary modified Tanimoto similarity which is unweighted ($\mathrm{MTU_{bin}}$)—in which each fingerprint has exactly one maximally dissimilar counterpart—by setting a constant $p = 0.5$, i.e.:

$$\mathrm{MTU_{bin}} = \frac{1}{2}(\mathrm{Tan_{bin}} + \mathrm{TanAbsence_{bin}}) \quad (6)$$

where similarity ranges from 0 (maximal dissimilarity) to 1 (identity), as above. The continuous version is given by:

$$\mathrm{MTU_{cont}} = \frac{1}{2}(\mathrm{Tan_{cont}} + \mathrm{TanAbsence_{bin}}) \quad (7)$$

The same modification is made to $\mathrm{MTW_{bin}}$ (see eq 4) to obtain a version which considers common presence in a continuous manner, i.e. $\mathrm{MTW_{cont}}$:

$$\mathrm{MTW_{cont}} = \frac{1}{3}(\mathrm{Tan_{cont}}(2 - p) + \mathrm{TanAbsence_{bin}}(1 + p)) \quad (8)$$

Hence, six similarity coefficients are considered in this work.

Because our Voronoi holograms are constructed by binning measured lengths and radii, a very small difference in these measurements can mean the difference between one bin and the next. To mitigate the effect of this bin-based thresholding, we diffuse (or smooth) the holograms prior to applying similarity coefficients. For each nonzero entry *q* in $h(v)$, we consider the (up to 26) neighboring points *r* in the cubic hologram representation and increase them by a distance-weighted fraction of $|q|$ (i.e., the magnitude of point *q*). Note that the unoccupied region of the grid representing $r_a < r_b$ remains unoccupied. The distance *d* is the Euclidean distance between the neighboring points in terms of steps; for example, a directly adjacent point has $d = 1$, whereas a point offset by one step in two axes has $d = \sqrt{2}$. The effect upon $|r|$ is given by

$$|r| = \frac{|q|}{d + 1} \quad (9)$$

For example, directly adjacent points *r* are increased in magnitude by $|q|/2$.

Implementation. The algorithms for generation of the described Voronoi holograms as well as tools for their comparison have been implemented in our Zeo++ software tool.[40] Zeo++ is a tool for performing high-throughput geometry-based analysis of porous materials and their voids. It also offers algorithms for the calculation of pore diameters, probe-accessible surface area, and volumes. Zeo++ is based on the Voro++ Voronoi library.[41]

■ DATASETS

Zeolites are microporous, crystalline materials comprised of periodically arranged SiO_4 tetrahedra. Approved zeolite framework types are catalogued in the IZA database; the version of the IZA database used in this study contains 185 zeolites after removing incomplete framework structures. From this set, we have selected 148 zeolites which have a channel system accessible to a spherical probe of 1.625 Å radii corresponding to a CH_4 molecule. We also analyze Deem's database of hypothetical zeolite frameworks; the Deem database used in this work consists of 331 171 zeolites (PCOD set),[16] from which we again select only CH_4-accessible structures, of which there are 139 397.

ZIFs are a related family of materials; they possess the same pore topology as zeolites, but their "building blocks" are zinc atoms and imidazole groups, resulting in larger periodic cells. Their close relationship to zeolites makes them ideal structures with which to further assess the performance of our diversity-based selection procedure. We generate ZIF frameworks computationally by substituting atoms in existing zeolite topologies; this method will be described in detail in a subsequent publication. Briefly, we substitute the Si−O−Si chain found in zeolites with the Zn−Im−Zn chain, where Im is the imidazole group, since the angles these chains form are similar.[5] Due to the different size of O and Im, the unit cells for ZIFs are rescaled in respect to their zeolite counterparts by a factor of 1.96. We apply this technique to the IZA zeolite set described above and again select only the CH_4-accessible structures. We find that all 185 IZA ZIFs are CH_4-accessible.

■ RESULTS

Profiling on IZA Zeolites and ZIFs. We have investigated and compared the considered similarity measures in application to Voronoi holograms. We compared the ability of Tanimoto, MTW, and MTU—in both binary and continuous forms—to select a diverse and representative sample of structures from the IZA zeolite database. We profiled the hologram density (number of active points in the hologram, i.e., the number which are nonzero) across the IZA set and also across 10 samples of the top 15 diverse structures retrieved by each similarity coefficient, with a different seed structure used in each sample. The aim was to determine whether the distribution of hologram densities within IZA could be reflected in a diverse, and hence representative, sampling.

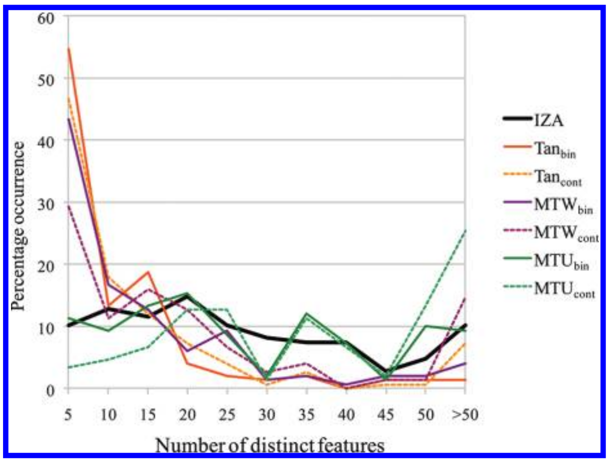


Figure 4. Profile of the number of distinct features in the IZA zeolite data set and in the structures retrieved in the top 15 hits in diversity-based selection using binary and continuous versions of three similarity coefficients. Percentages are averaged across 10 samples with different seeds (same 10 seeds used for each coefficient).

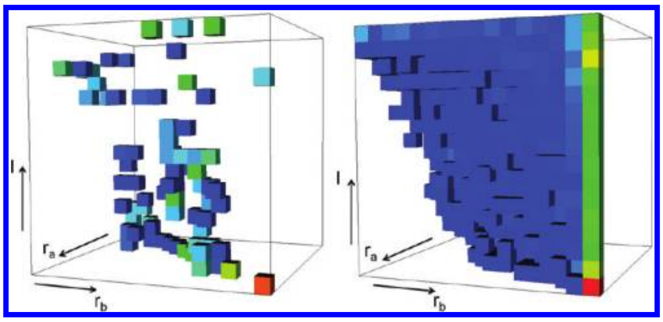


Figure 5. Left: the Voronoi hologram for ZIF SOD, the sparsest in the IZA ZIF set. The frequency of feature occurrence ranges in this case from 2 (dark blue) to 132 (red, obscured). Right: the Voronoi hologram for ZIF IMF, the densest in the IZA ZIF set. The frequency of feature occurrence ranges in this case from 2 (dark blue) to 1072 (red).

Figure 4 illustrates that the Tanimoto coefficient exhibits the documented bias toward simple (i.e., sparse hologram) structures; however MTW also exhibits a subtle bias. The MTU coefficient gives a more representative sampling of hologram complexity, and so we proceed with this coefficient. Comparing the binary and continuous versions of these coefficients, we find subtle differences. The continuous versions demonstrate retrieval of a smaller quantity of sparse holograms, and a larger quantity of dense holograms. This is intuitive, as points set in both arrays may have varying multiplicities, in which case the similarity is reduced compared to a binary similarity; hence, continuous coefficients will tend to select more dense structures.

As well as zeolites, we are also interested in characterization of ZIFs, a class of MOF. As described above, ZIFs are more complex structures than zeolites, since each oxygen atom in a zeolite corresponds to an imidazole ring in its ZIF counterpart. ZIF periodic unit cells are hence larger due to the size of the imidazole ring being substituted for the zeolite oxygen atoms. While the densest hologram in IZA zeolites, TUN, has 145 distinct features present (see), the sparsest hologram in IZA ZIFs, SOD, has 132, and the densest, IMF, has 1098 (see Figure 5). The average number of distinct features is 27.155 for zeolites and 549.130 for ZIFs. Therefore, it is interesting to examine how Voronoi holograms behave when applied to this different class of material.

As for IZA zeolites, we profile the performance of our six similarity coefficients with respect to IZA ZIFs (see Figure 6). Although the difference in behavior between these two classes of material is pronounced, we still find that binary coefficients have a stronger propensity to retrieve simple structures. Within either the binary or continuous coefficients, Tanimoto, MTW, and MTU perform similarly.

We observe, therefore, that while diversity-based selection using MTU_{bin} provides a representative spread of hologram density for IZA zeolites, as the density of structures increases significantly, all binary coefficients begin to exhibit Tanimoto's known bias toward sparsity. For very dense representations, therefore, a continuous coefficient is most appropriate.

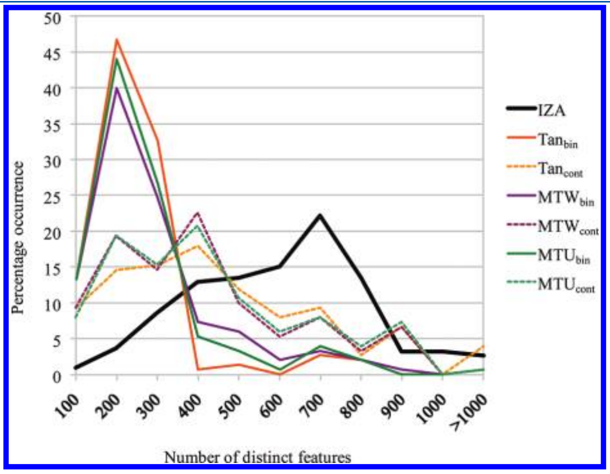


Figure 6. Profile of the number of distinct features in the IZA ZIF data set and in the structures retrieved in the top 15 hits in diversity-based selection using binary and continuous versions of three similarity coefficients. Percentages are averaged across 10 samples with different seeds (same 10 seeds used for each coefficient).

Diverse IZA Zeolites. Following this validation experiment, we aimed to select a single representative subset of zeolite structures in IZA. We arbitrarily selected the alphabetically first structure, ABW, as the seed. By observation, we determined that an MTU_{bin} similarity threshold of 0.5, at which point a total of 20 structures are present in the sample, constitutes an intuitive position at which to terminate selection. This subset consists of a range of visually distinct features: narrow and wide channels, junctions of various sizes and degrees of connectivity, large and small cages, and pronounced and subtle widening and narrowing of accessible regions. Figure 7 demonstrates this with pore network contours; the highlighted regions of the structures are the pore networks, lighter regions being the inside of pores and darker regions the outside. For instance, the first structure, ABW, consists of two very thin channels which each cross the periodic boundary. Certain structures in the diverse set appear similar—however, their selection is the result of differences in the hologram features set in each structure, which reflect more subtle differences in their overall shapes. For instance, EAB and AFT are somewhat similar (0.48 MTU_{bin} similarity); however, the differences in the shapes of their pore network cages and connecting channels result in their differing hologram densities: they exhibit

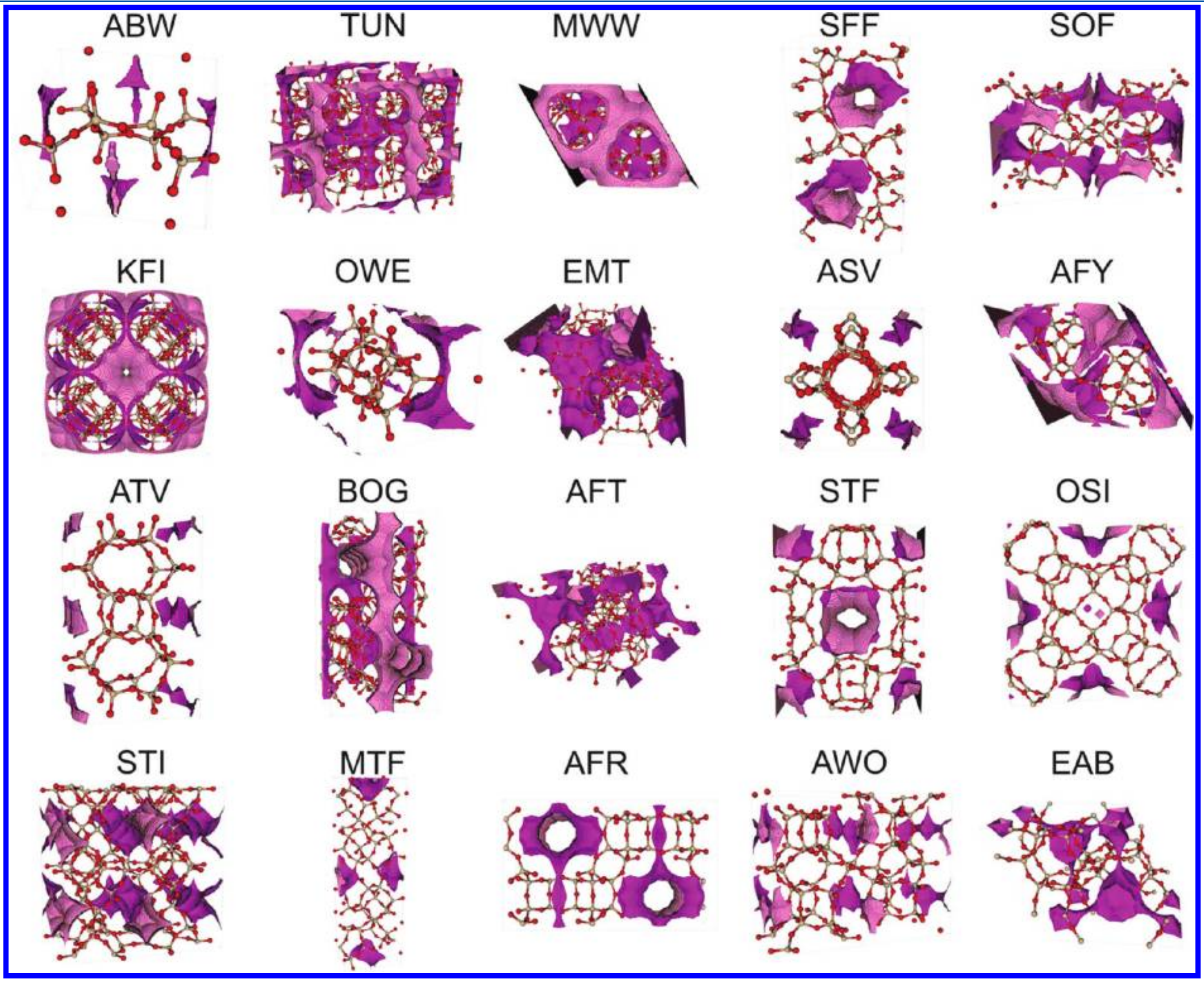


Figure 7. The pore networks for the first 20 (before 0.5 similarity threshold) IZA structures selected by the diversity-based selection method.

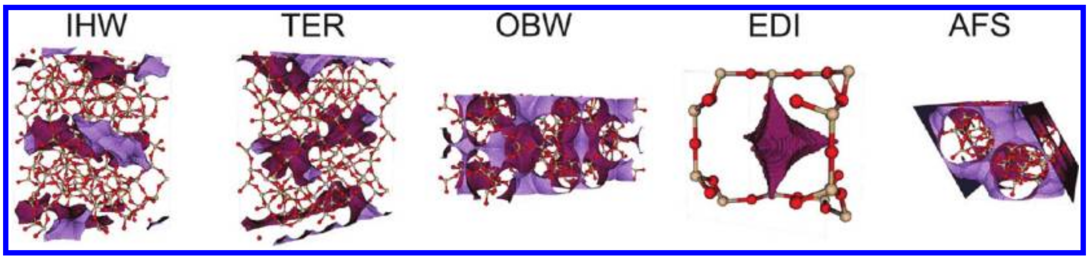


Figure 8. The pore networks for the following five (after 0.5 similarity threshold) IZA structures retrieved by the diversity-based selection method.

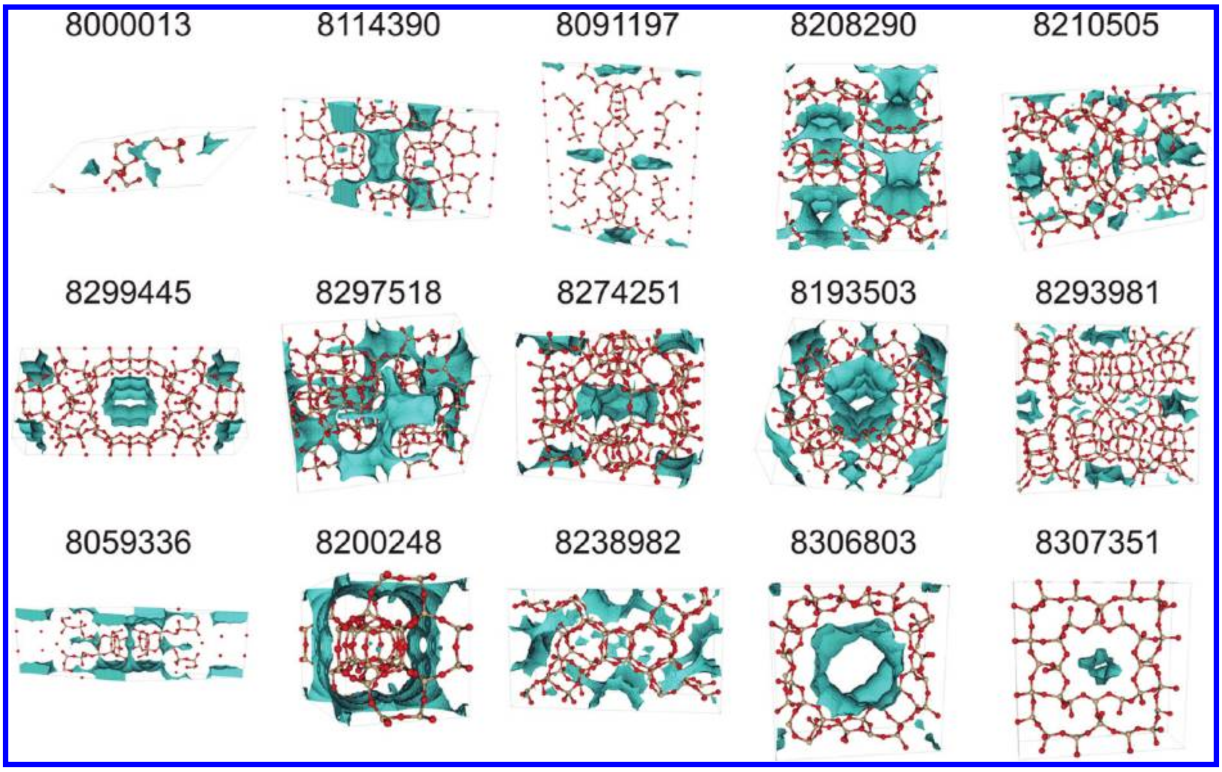


Figure 9. The pore networks for the first 15 hypothetical structures selected by the diversity-based selection method.

16 and 31 distinct features respectively; EAB's smaller cages (d_i = 7.08 Å) exhibit less diversity in features present than AFT (d_i = 7.69 Å).

Continuing to select structures beyond this point retrieves those with greater than 0.5 similarity to a previously selected structure, and by inspection some similarity between candidates and previously selected structures begins to emerge (see Figure 8). For instance, structures MWW and AFY in Figure 7 and AFS in Figure 8 appear to be similar, each with a 120° inclined parallelepiped unit cell and clear cages centered on the corners; however, MWW and AFY are retrieved in our selection of the top 20 structures and AFS is not. AFS exhibits a free sphere diameter of 5.95 Å, similar to AFY's 5.84 Å (MWW has 4.86 Å), while it has an included sphere diameter of 9.45 Å, similar to MWW's 9.63 Å (AFY has 7.76 Å). As seen for EAB and AFT, the difference in the basic structural properties observed here contributes directly toward dissimilarity due to the activation of different features in the structures' holograms. MWW and AFY are found to be diverse because of a lack of features in common—this is visible in the near-constant channel and cage diameter in MWW, which by contrast varies highly in AFY, as in AFS.

Diverse Hypothetical Zeolites. Following our analysis of the IZA set, we apply the same approach to select a diverse subset of hypothetical zeolites. We seed the method with the numerically first structure, with ID 8000013, and retrieve a total of 174 structures before reaching the 0.5 MTU_{bin} threshold. We provide pore network diagrams of the most diverse 15 (i.e., those retrieved first) in Figure 9. The entire list of structures in this diverse set is included in the Supporting Information. It is clear that as observed for the IZA set, our diverse sampling of hypothetical zeolites retrieves structures with a range of visually distinct features; for instance, there is a mixture of narrow and wide channels. However, what is quite striking about this selection, and which we do not find for IZA zeolites, is the high percentage of one-dimensional channel systems (71.8%, with 18.4% and 9.8% two- and three-dimensional, respectively). We find however that this is broadly representative of the hypothetical set at large, of which 87.3% are one-dimensional channels (with 8.2% and 4.5% two- and three-dimensional respectively). Moreover, the range of dimensionalities retrieved by our method is not what would be expected with a random selection (which would tend to select percentages equal to those observed in the

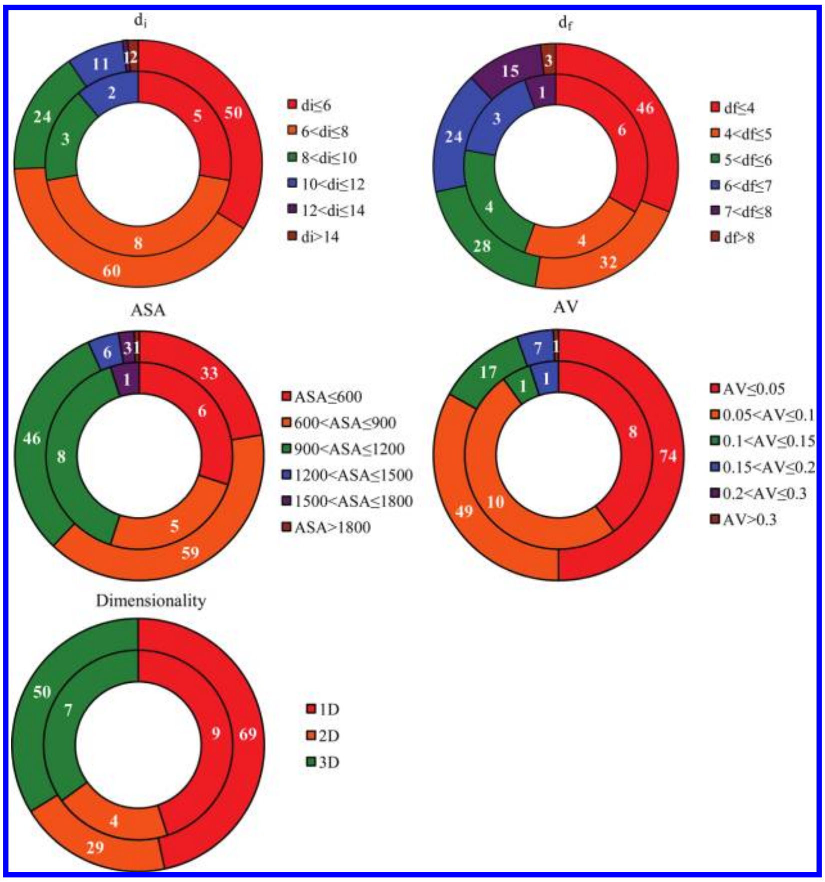


Figure 10. Pie charts illustrating the range of the five basic structural properties for the CH_4-accessible IZA set (outer ring) and the 20 most diverse zeolites retrieved using the MTU_{bin} coefficient applied to Voronoi holograms (inner ring).

whole set), nor a selection explicitly based upon dimensionality (which would prioritize outliers). Rather, our method provides a diverse and representative sample, maintaining the majority of one-dimensional channels while slightly increasing the proportion of two- and three-dimensional channels. This characteristic of our method is discussed further in the following section.

■ DISCUSSION

We have demonstrated that Voronoi holograms facilitate the selection of zeolite structures with geometrically diverse pores. We focus on IZA zeolites using MTU_{bin} as described above. Two questions remain: (1) Are the diverse structures representative? That is, do they give an overview of the variety present in the data set? (2) Can the same behavior be achieved through the use of the existing structural descriptors, namely, free and included sphere diameter, accessible surface area and volume, or dimensionality? That is, do Voronoi holograms provide a qualitatively new means of comparing structures?

First, we explore the representative quality of the diverse 20 IZA zeolites, i.e., those obtained before a 0.5 MTU_{bin} threshold is reached. We plot in Figure 10 the observed distributions of basic structural properties in both IZA and the diverse sample. We find that the diversity-based selection constitutes a representative sample of the IZA data set with respect to pore network dimensionality and to varying degrees also maps the observable range of the remaining basic structural properties. We note in particular that for these other four properties, the structures within IZA exhibiting the largest values are not chosen by our method. This puts our method into stark contrast with a diversity-based selection performed upon either an individual basic structural descriptor or a group thereof, as described below.

A diversity-based selection, which uses one or more of these simple descriptors, will invariably prioritize the selection of a set of extreme outlier structures, which exhibit the most diverse range of the basic structural properties considered. For instance, using d_i, it is clear that the first few structures selected will have included sphere diameters from the periphery of observed values, and subsequent selections will be distributed as evenly as possible through this property space irrespective of the distribution of included sphere diameters observed within the data set. If one desires a set of structures with the most diverse range of a specific property, then this is acceptable; however, this process reveals nothing about the distribution of this or any other property throughout the data set and considers no other aspect of the structures in question (for instance, one might retrieve a diverse range of included sphere diameters but a very narrow range of free sphere diameters). This problem can be mitigated by the fusion of similarities calculated with respect to a range of structural descriptors. For instance one might define the similarity between two structures as a fusion of the similarity of their included sphere and their free sphere diameters; yet, with such a two-dimensional similarity, it will invariably be the case that structures from the extreme corners of this two-dimensional property space are selected, after which an even distribution will arise, but still the problem remains that this selection does not

Table 2. Comparison of the Pearson's Correlation between Each of the Descriptors Described above to Three Decimal Places[a]

Pearson's correlation between similarity measures	d_i	d_f	ASA	AV	dimensionality	holograms
d_i	1.000	0.313	0.271	0.672	0.076	0.332
d_f	0.313	1.000	0.011	0.225	0.009	0.254
ASA	0.271	0.011	1.000	0.618	0.271	0.115
AV	0.672	0.225	0.618	1.000	0.125	0.215
dimensionality	0.076	0.009	0.271	0.125	1.000	0.102
holograms	0.332	0.254	0.115	0.215	0.102	1.000

[a] ASA, AV, dimensionality, and holograms are with respect to a CH_4 probe.

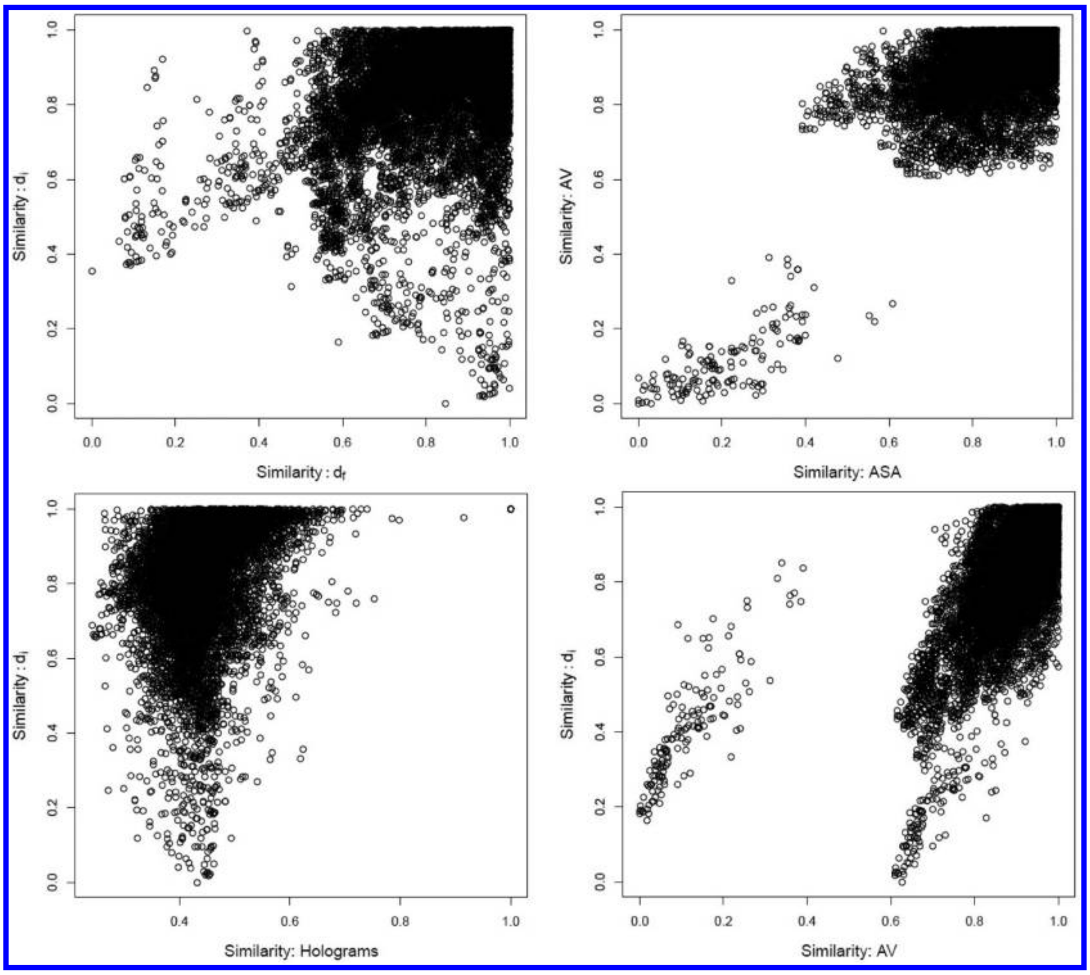


Figure 11. Plots illustrating how structural similarities measured using basic structural descriptors, as well as holograms using MTU_{bin} similarity, are related. All pairwise comparisons for 148 IZA zeolites are plotted.

reveal any information about the distribution of properties, and it is an "artificial", deliberate selection biased toward some favored structural properties. These selections may be diverse with respect to the chosen properties, but they will not be "representative" of the data set at large.

Hence, a comparison routine based upon a fusion of similarities from basic structural descriptors will be useful only if the selection of a diverse range of these properties is the only concern. We demonstrate that Voronoi holograms provide a qualitatively different means of selecting structures by investigating the degree to which correlation exists between the five basic structural descriptors as well as holograms. For each pair of structures in the 148-member IZA set, their similarity with respect to each descriptor is calculated; for the one-dimensional basic structural descriptors, similarity is defined as the normalized distance between the measured properties, such that the two most extreme values observed yield a similarity of zero (for dimensionality, similarity is binary: 1 if the structures have the same dimensionality, else 0). The result is a matrix of pairwise similarities for each descriptor, each row (or column) of which gives the observed similarities between some query structure and each other structure in the data set. Matrices are compared by calculating the average Pearson product-moment correlation coefficient[42] between the rows (equivalently, columns) of the matrices, and the resulting correlations between matrices are provided in Table 2, with Figure 11 illustrating some of the most interesting comparisons. The units of d_i and d_f used in this analysis are Ångströms, and the units of ASA and AV are square meters per gram and cubic centimeters per gram, respectively. The strongest correlation between two different descriptors is found in AV (for a CH_4 probe approximated with a sphere of radii 1.625 Å) and d_i; this is intuitive, since a large included sphere tends to indicate a large accessible volume—however, d_i does not reveal the quantity of large cages, the variance in which will reduce this correlation. The only other high correlation observed is between AV and ASA; this is also intuitive, since a large accessible volume will tend to have a large surface area—however, very thin channels will exhibit a larger surface area than a cage of the same volume, reducing the correlation. Of the remaining descriptors, dimensionality is the least highly correlated with any other, although this is unsurprising given the binary nature of dimensionality comparison. Finally, d_f and holograms are not highly correlated with any other existing basic structural descriptors, or with each other; their highest correlations by magnitude are with d_i. However, we duly note that a lack of correlation alone is not indicative of the presence of valuable structural information absent in other descriptors; for instance, one might construct a novel descriptor uncorrelated with any known measure which is based on the three letter names assigned to each IZA structure.

Nevertheless, we argue that Voronoi holograms are an improvement upon these basic structural descriptors. Voronoi holograms are higher-dimensional descriptors (with 2176 degrees of freedom), in contrast to the one-dimensional descriptors described above; as such, diversity selection will move toward selecting structures from the periphery of this higher-dimensional property space, which necessarily involves the selection of structures which possess differing features (as defined above, i.e., edges connecting Voronoi nodes of specific radii). A single structure exhibits many such features, whereas it will for instance exhibit only a single included sphere diameter. Each hologram feature describes the shape or "texture" of some localized part of a pore network, reflecting the arrangement of local atoms, whereas for instance the included sphere diameter describes the entire pore network with a single number. We contend that our holograms, through their abstraction of the entire Voronoi network and comparative complexity, constitute a representation of porous structures, which is more revealing of the overall shape of a pore network.

Finally, another interesting question is whether the geometrically diverse sets of materials selected by the presented method will also have diverse physical properties. Although this issue will be investigated in detail in our future studies, other preliminary results suggest that if the physical property depends on the entire structure and its void space rather than a specific local feature, the diverse set will likely present a wide range of this property. For example, in our recent study, we have used multiscale modeling to predict the energy required to capture CO_2 from flue gases of a power plant (referred to as the parasitic energy) using adsorption-based separations and a porous material.[43] Our results suggest that our diverse set of materials cover most of the range of parasitic energies. At the same time, we observed that other physical properties such as the Henry coefficient for CO_2 can be practically determined by a local arrangement of atoms forming a preferential adsorption site, and therefore diversity selection may not pick up structures with such sites and exceptionally high Henry coefficients.

CONCLUSIONS

The development of large databases of porous materials has to trigger the development of cheminformatics tools to analyze, select, group, and classify those materials. We have demonstrated an approach that allows efficient diversity selection of structures based on geometrical and shape characteristics of materials. Our approach employs Voronoi decomposition as a technique to convert material structure into a periodic graph representation of the material's void space. Then, the guest molecule accessible fragment of the Voronoi network can be used to obtain a hologram representation, which in turn can be compared using a modified Tanimoto coefficient.

Our investigation suggests that the proposed approach provides an essentially new way to compare structures on the basis of pore characteristics, as our similarity does not correlate with similarity defined on the basis of other recently proposed structural descriptors such as free sphere diameter and largest included sphere, as well as commonly used descriptors such as accessible volume and surface area.

ASSOCIATED CONTENT

S Supporting Information. A list of IDs of the most diverse hypothetical zeolites identified in this work is provided with this article. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Fax: (510) 486 58 12. E-mail: mharanczyk@lbl.gov.

ACKNOWLEDGMENT

The authors would like thank Michael W. Deem for providing the database of hypothetical zeolites as well as Chris H. Rycroft, Thomas Willems, and Peter Willett for comments on the presented approach. This work was supported in part (to R.L.M. and M.H.) by the U.S. Department of Energy under contract DE-AC02-05CH11231 and through SciDAC project #CSNEW918 entitled "Knowledge guided screening tools for identification of porous materials for CO_2 separations". B.S. was supported as part of 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.

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.

REFERENCES

(1) *Handbook of Zeolite Science and Technology*; Auerbach, S. M., Carrado, K. A., Dutta, P. K., Ed.; Marcel Dekker: New York, 2004.

(2) Smit, B.; Maesen, T. L. M. Towards a molecular understanding of shape selectivity. *Nature* **2008**, *457*, 671–677.

(3) Smit, B.; Maesen, T. L. M. Molecular Simulations of Zeolites: Adsorption, Diffusion, and Shape Selectivity. *Chem. Rev.* **2008**, *108*, 4125–4184.

(4) Krishna, R.; van Baten, J. M. Using molecular simulations for screening of zeolites for separation of CO2/CH4 mixtures. *Chem. Eng. J.* **2007**, *133*, 121–131.

(5) Phan, A.; Doonan, C. J.; Uribe-Romo, F. J.; Knobler, C. B.; O'Keeffe, M.; Yaghi, O. M. Synthesis, structure, and carbon dioxide capture properties of Zeolitic Imidazolate Frameworks. *Acc. Chem. Res.* **2010**, *43*, 58–67.

(6) Millward, A. R.; Yaghi, O. M. Metal-organic frameworks with exceptionally high capacity for storage of carbon dioxide at room temperature. *J. Am. Chem. Soc.* **2005**, *127*, 17998–17999.

(7) Walton, K. S.; Millward, A. R.; Dubbeldam, D.; Frost, H.; Low, J. J.; Yaghi, O. M.; Snurr, R. Q. Understanding inflections and steps in carbon dioxide adsorption isotherms in metal-organic frameworks. *J. Am. Chem. Soc.* **2008**, *130*, 406–407.

(8) Banerjee, R.; Phan, A.; Wang, B.; Knobler, C.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. High-throughput synthesis of zeolitic imidazolate frameworks and application to CO2 capture. *Science* **2008**, *319*, 939–943.

(9) Sumida, K.; Hill, M. R.; Horike, S.; Dailly, A.; Long, J. R. Synthesis and Hydrogen Storage Properties of Be-12(OH)(12)(1,3, 5-benzenetribenzoate)(4). *J. Am. Chem. Soc.* **2009**, *131*, 15120–15121.

(10) Choi, H. J.; Dinca, M.; Long, J. R. Broadly hysteretic H-2 adsorption in the microporous metal-organic framework Co(1, 4-benzenedipyrazolate). *J. Am. Chem. Soc.* **2008**, *130*, 7848–7450.

(11) D'Alessandro, D. M.; Smit, B.; Long, J. R. Carbon Dioxide Capture: Prospects for New Materials. *Angew. Chem., Int. Ed.* **2010**, *49* (35), 6058–6082.

(12) (a) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E.; Li, X.; Karen, V. L. Framework-Type Determination for Zeolite Structures in the Inorganic Crystal Structure Database. *J. Phys. Chem. Ref. Data* **2010**, *39*, 033102–033145. (b) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. Identifying Zeolite Frameworks with a Machine Learning Approach. *J. Phys. Chem. C* **2009**, *113*, 21721–21725.

(13) Foster, M. D.; Treacy, M. M. J. http://www.hypotheticalzeolites.net (accessed Nov 13, 2009).

(14) Earl, D. J.; Deem, M. W. Toward a Database of Hypothetical Zeolite Structures. *Ind. Eng. Chem.* **2006**, *45*, 5449–5454.

(15) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* **2009**, *113*, 21353–21360.

(16) Pophale, R.; Cheeseman, P. A.; Deem, M. W. A Database of New Zeolite-Like Materials. *Phys. Chem. Chem. Phys.* **2011**, *13*, 12407–12412.

(17) Haldoupis, E.; Nair, S.; Sholl, D. S. Efficient Calculation of Diffusion Limitations in Metal Organic Framework Materials: A Tool for Identifying Materials for Kinetic Separations. *J. Am. Chem. Soc.* **2010**, *132*, 7528–7539.

(18) Blatov, V. A.; Delgado-Friedrichs, O.; O'Keeffe, M.; Proserpio, D. M. Three-periodic nets and tilings: natural tilings for nets. *Acta Crystallogr.* **2007**, *A63*, 418–425.

(19) (a) Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Novel Approach for Clustering Zeolite Crystal Structures. *Mol. Inform.* **2010**, *29*, 297–301. (b) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine learning approach for structure-based zeolite classification. *Microporous Mesoporous Mater.* **2009**, *117*, 339–349.

(20) Düren, T.; Millange, F.; Férey, G.; Walton, K. S.; Snurr, R. Q. Calculating Geometric Surface Areas as a Characterization Tool for Metal−Organic Frameworks. *J. Phys. Chem. C* **2007**, *111*, 15350–15356.

(21) Foster, M. D.; Rivin, I.; Treacy, M. M. J.; Delgado, O. A geometric solution to the largest-free-sphere problem in zeolite frameworks. *Microporous Mesoporous Mater.* **2006**, *90*, 32–38.

(22) Li, H.; Laine, A.; O'Keeffe, M.; Yaghi, O. M. Supertetrahedral Sulfide Crystals with Giant Cavities and Channels. *Science* **1999**, *283*, 1145–1147.

(23) Haldoupis, E.; Nair, S.; Sholl, D. S. Pore size analysis of >250000 hypothetical zeolites. *Phys. Chem. Chem. Phys.* **2011**, *13*, 5053–5060.

(24) First, E. L.; Gounaris, C. E.; Wei, J.; Floudas, C. A. Computational characterization of zeolite porous networks: an automatic approach. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17339–17358.

(25) Wei, J.; Floudas, C. A.; Gounaris, C. E.; Somorjai, G.A.. Search engines for shape selectivity. *Catal. Lett..* **2009**, *133*, 234–241.

(26) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134–141.

(27) Haranczyk, M.; Sethian, J. A. Navigating molecular worms inside chemical labyrinths. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 21472–21477.

(28) Haranczyk, M.; Sethian, J. A. Automatic structure analysis in high-throughput characterization of porous materials. *J. Chem. Theory Comput.* **2010**, *6*, 3472–3480.

(29) Theisen, K.; Smit, B.; Haranczyk, M. Chemical Hieroglyphs: Abstract Depiction of Complex Void Space Topology of Nanoporous Materials. *J. Chem. Inf. Model.* **2010**, *50*, 461–469.

(30) Beketayev, K.; Weber, G. H.; Haranczyk, M.; Bremer, P.-T.; Hlawitschka, M.; Hamann, B. Topology-based Visualization of Transformation Pathways in Complex Chemical Systems. *Comput. Graphics Forum* **2011**, *30*, 663–673.

(31) Blatov, V. A. Voronoi-Dirichlet polyhedra in crystal chemistry: theory and applications. *Cryst. Rev.* **2004**, *10*, 249–318.

(32) Blatov, V. A.; Shevchenko, A. P. Analysis of voids in crystal structures: the methods of 'dual' crystal chemistry. *Acta Crystallogr.* **2003**, *A59*, 34–44.

(33) Alinchenko, M. G.; Anikeenko, A. V.; Medvedev, N. N.; Voloshin, V. P.; Mezei, M.; Jedlovszky, P. Morphology of Voids in Molecular Systems. A Voronoi−Delaunay Analysis of a Simulated DMPC Membrane. *J. Phys. Chem. B* **2004**, *108*, 19056–19067.

(34) Blatov, V. A.; Ilyushin, G. D.; Blatova, O. A.; Anurova, N. A.; Ivanov-Schits, A. K.; Dem'yanets, L. N. Analysis of migration paths in fast-ion conductors with Voronoi-Dirichlet partition. *Acta Crystallogr.* **2006**, *B62*, 1010–1018.

(35) Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik* **1959**, *1*, 269–271.

(36) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Modell.* **1997**, *15*, 372–385.

(37) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, the Netherlands, 2003; pp 103−104, 134−136.

(38) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 379–386.

(39) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44*, 110–119.

(40) http://www.carboncapturematerials.org/Zeo++/ (accessed 11/2/11). The source code is available from the authors upon request.

(41) http://math.lbl.gov/voro++/ (accessed 11/2/11).

(42) Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. R. Soc. London, Ser. A* **1896**, *187*, 253–318.

(43) Lin, L.-C.; Berger, A.; Martin, R. L.; Kim, J.; Swisher, J.; Jariwala, K.; Rycroft, C. H.; Bhown, A.; Deem, M. W.; Haranczyk, M.; Smit, B. In silico screening of carbon capture materials. *Nat. Mater.* Submitted.