Journal of Molecular Graphics and Modelling 44 (2013) 208–219

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Characterization and comparison of pore landscapes in crystalline porous materials

Marielle Pinheiro [a], Richard L. Martin [a], Chris H. Rycroft [a,b], Andrew Jones [c], Enrique Iglesia [c,d], Maciej Haranczyk [a,*]

[a] Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720-8139, USA
[b] Department of Mathematics, University of California, Berkeley, CA 94720, USA
[c] Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA
[d] Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 6 March 2013
Received in revised form 3 May 2013
Accepted 11 May 2013
Available online 17 June 2013

Keywords:
Porous materials
Pore size distribution
Stochastic rays
Pore shape similarity

ABSTRACT

Crystalline porous materials have many applications, including catalysis and separations. Identifying suitable materials for a given application can be achieved by screening material databases. Such a screening requires automated high-throughput analysis tools that characterize and represent pore landscapes with descriptors, which can be compared using similarity measures in order to select, group and classify materials. Here, we discuss algorithms for the calculation of two types of pore landscape descriptors: pore size distributions and stochastic rays. These descriptors provide histogram representations that encode the geometrical properties of pore landscapes. Their calculation involves the Voronoi decomposition as a technique to map and characterize accessible void space inside porous materials. Moreover, we demonstrate pore landscape comparisons for materials from the International Zeolite Association (IZA) database of zeolite frameworks, and illustrate how the choice of pore descriptor and similarity measure affects the perspective of material similarity exhibiting a particular emphasis and sensitivity to certain aspects of structures.

1. Introduction

Crystalline porous materials exhibit complex networks of internal void space, which can permit the adsorption and diffusion of guest species. The shape, size and chemistry of localized regions within these networks determine the interactions that can occur between a guest and the host material. Many applications in the chemical industry take advantage of the pore properties of these materials. For example, zeolites are particularly important as cracking catalysts in oil refining [1], and they also find application as alkylation and isomerization catalysts and materials for separations [2–5]. In addition to zeolites, metal organic frameworks (MOFs) [6,7] and zeolitic imidazolate frameworks (ZIFs) [8] have generated interest for their potential use in gas separation and storage [9–11]. The search for materials with better performance is ongoing. Roughly 194 zeolites with 1400 various chemical compositions have been synthesized [12], while millions of new, computationally generated zeolites await investigation [13–16]. Thousands of MOFs have been synthesized in the last decades [17], and large databases of hypothetical structures are being compiled [18].

Databases of materials can, in principle, be screened to identify materials with useful and superior catalytic, storage or separation properties. The current state-of-the-art computational methodology, based on molecular simulations and electronic structure calculations, can be used to accurately predict adsorption and/or catalytic properties of a particular structure and provide reliable information on its performance in specific applications, such as hydrodewaxing [3] and CO_2 separation [19,20]. However, finding the optimal material for a given application remains a formidable challenge because the number of possible materials is extremely large. Fortunately, recent advances in high-performance parallel computing and void space analysis algorithms have enabled simulation of material sets as large as 100,000 structures [18,20]. At the same time, the development of large material databases and high-throughput molecular simulations have brought new challenges related to material data analysis. New, automated computational and cheminformatics techniques need to be developed to characterize, categorize, and search such large databases as well as enable data-mining for useful knowledge.

A key feature of automated material analysis tools is that they convert structural information describing a material, *e.g.* positions

* Corresponding author. Tel.: +1 510 486 7749; fax: +1 510 486 5812.
E-mail address: mharanczyk@lbl.gov (M. Haranczyk).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.05.007

of atoms in a periodic unit cell, into one or more sets of descriptors. Descriptors can be symbols, letter codes or numbers that capture important features of a material and thus enable structure comparison. Recently, a number of algorithms and software tools have been developed to allow automatic analysis of a material's structure and its void space. For example, Foster et al. and Haldoupis et al. have presented methods to calculate geometrical parameters describing pores in zeolite materials [21,22], namely the diameter of the largest included (D_i) and the largest free (D_f) spheres [23]. The largest included sphere reflects the size of the largest cavity within a porous material, whereas the largest free sphere corresponds to the largest spherical probe that can diffuse fully through a structure and measures a minimum restricting aperture along a diffusion path. Other descriptors include accessible volume (AV) and accessible surface area (ASA) with efficient Monte Carlo algorithms for calculation thereof explored by Düren et al. [24,25] and Do et al. [26].

Another approach to the analysis of porous materials is a fragment-based approach where the basic building units for a material are identified and characterized. For example, Blatov and coworkers have pursued the concept of natural tiling of periodic networks to find and classify primitive building blocks in zeolites [27]. First et al. have developed ZEOMICS and MOFOMICS tools [28] that fragment and represent the void space inside porous materials as a set of geometrical blocks such as cylinders and spheres. Another approach, which combines simple structure representations and local information about the void space, is characterization by pore size distribution (PSD) histograms. PSD histograms indicate the fraction of the void space volume that corresponds to certain pore sizes. They lack, however, information about pore connectivity. Algorithms for PSD and their implementations were recently discussed by Do et al. [26] and Sarkisov et al. [29], and involved computational geometry and grid-based approaches, respectively.

Our group has also contributed algorithms and software for high-throughput geometry-based analysis of materials and their void space [30–32]. The core of our approaches is a computational geometry technique, the Voronoi decomposition [33]. In the Voronoi decomposition, the space surrounding n atoms is divided into n polyhedral cells such that each cell face is a plane equidistant from the two atoms sharing that face. Edges of such cells overlap with lines equidistant to three neighboring atoms, whereas vertices of cells, the Voronoi nodes, are equidistant from four neighboring atoms (in a general asymmetric case). The Voronoi network, a three-dimensional graph comprising such nodes and edges, maps the void space surrounding the atoms. Analysis of such a network can provide parameters such as D_i and D_f and detailed information about void space geometry and topology [32]. For example, void space regions inaccessible to a given probe can be identified, and this information can be utilized in the calculation of ASA and AV [32]. Moreover, the Voronoi decomposition served as a starting point for the development of a novel structure descriptor, which encodes the entire shape of the void space for a given material: the Voronoi hologram [34]. The Voronoi hologram is a histogram representation of the guest-accessible portion of the Voronoi network. Its development allowed us to establish a framework for efficient navigation through the chemical space of all-siliceous zeolites [34]. Our algorithms have been implemented in the Zeo++ package [35] and use the Voro++ software library [36,37]. The Voronoi decomposition has also been used by others in the analysis of crystalline materials [38] and their voids [39], as well as membranes [40] and has been suggested as a tool to investigate ion transport pathways in crystals [41].

In the current contribution, we present new algorithms that allow high-throughput automatic analysis of pore landscapes of porous materials and generation of the corresponding descriptors. In particular, we demonstrate an efficient algorithm for PSD calculation as well as highlight recently developed stochastic ray-tracing histograms [42]. Finally, we illustrate and discuss applications of these and previously developed descriptors in calculating pore-shape similarities, which then can be used in database searching and diverse structure sampling.

2. Methods

2.1. Calculation of the Voronoi network and prediction of void space accessibility

The calculation of the Voronoi network and its analysis to predict void space accessibility have been described in detail in Ref. [32] and will therefore only be briefly introduced here. The Voronoi network is computed using a modified version of Voro++, an open source library for three-dimensional Voronoi calculations [36,37]. The library is based upon individually computing the Voronoi cell associated with each atom, which is stored internally as a collection of edges and vertices. During computation of the Voronoi network, several important parameters are tabulated for use in later analysis. For example, for each edge and node (vertex), the minimum distance to an atom (or its surface) is stored and we denote these as node radii. The library can also handle chemical systems with atoms of various radii by making use of the radical Voronoi tessellation, which is a common approach to handle polydisperse particle arrangements and has been previously shown to be a good method of constructing a network for porosity calculations of unequally sized spheres [43].

The obtained Voronoi network represents the void space in a porous material. Analysis of such a Voronoi network using Dijkstra-like graph algorithms [44] allows the detection of probe-accessible subnetworks. This information can then be used to determine if any point inside a periodic unit cell is accessible to a probe. Here, an accessible point is defined as a point that can be reached by the center of a spherical probe of a given radius. In practice, the Voronoi decomposition is performed on a copy of the atom network in which all atom radii are increased by an amount equivalent to the probe radius. Therefore, any point outside of such expanded particles can accommodate the spherical probe. The procedure for determining accessibility of points is described in detail in Ref. [32], and was employed in Monte Carlo sampling of accessible volume. The same procedure provides the basis for new algorithms presented herein.

2.2. Pore landscape characterization

2.2.1. Pore size distribution

The pore size distribution (PSD) provides information about the fraction of void space that is occupied by pores of certain size. The PSD can be calculated using a Monte Carlo approach similar to accessible volume calculations [32], but with each sampled point described by both its accessibility and pore size. The key features of our algorithm focus on the determination of the largest sphere that encapsulates the sampled point without overlapping any atoms of the structure. In the discussion below, the atoms are treated as having zero radii and the regular Voronoi tessellation is used. The generalization to take into account atom radii is presented later.

Let the atoms in a given structure have positions $\mathbf{x}_i$ for $i = 1, \ldots, n$ and let $d(\mathbf{x},\mathbf{y})$ be the Euclidean distance between two positions $\mathbf{x}$ and $\mathbf{y}$. For a given atom i, the Voronoi cell is defined to be

$$d(\mathbf{x}, \mathbf{x}_i) < d(\mathbf{x}, \mathbf{x}_j)$$

for all j not equal to i. The Voronoi cells form irregular polyhedra that tessellate the space. Here, we consider a general asymmetric case, wherein each Voronoi cell face is equidistant to two atoms, each

cell edge is equidistant to three atoms, and each Voronoi cell vertex is equidistant to four. In some cases, additional symmetries of the atomic arrangement may increase the number of atoms equidistant to faces, edges, and vertices, but these cases do not need to be considered in describing the PSD algorithm since they can be made into an asymmetric arrangement by moving atoms by arbitrarily small displacements.

At a given position **s**, the maximum radius of a pore centered at that location is given by the minimum distance to an atom,

$$R(\mathbf{s}) = \min\{d(\mathbf{s}, \mathbf{x}_i) : i = 1, \ldots, n\}.$$

This function consists of patches of the form d(**s**,$\mathbf{x}_i$) within each Voronoi cell, which connect continuously to neighboring patches. Now consider a sample point **p**: if a pore located at **s** covers **p**, then $d(\mathbf{s},\mathbf{p}) < R(\mathbf{s})$ and thus

$$d(\mathbf{s}, \mathbf{p}) < d(\mathbf{s}, \mathbf{x}_i)$$

for all *i*. This can be interpreted geometrically: the point **s** must be in the Voronoi cell $V_\mathbf{p}$ of **p** with respect to the extended set of points $\{\mathbf{p},\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$. We refer to $V_\mathbf{p}$ as the *ghost cell* of **p**, and it can be calculated with a straightforward extension of the Voronoi software library that makes use of the existing Voronoi cell computation routines. Hence, finding the largest pore containing **p** is equivalent to the constrained optimization problem of maximizing $R(\mathbf{s})$ within the volume $V_\mathbf{p}$.

To solve this, the algorithm must locate all local maxima of $R(\mathbf{s})$ within $V_\mathbf{p}$ and choose the largest. There are several possible cases. First, a maximum may lie in the interior of $V_\mathbf{p}$, in which case it will be at a node of a Voronoi network – these points can be tested quickly from the previously stored network information.

Otherwise, a local maximum must lie on the boundary of $V_\mathbf{p}$. The first possibility is for a maximum to lie in the interior of a face of $V_\mathbf{p}$. However, the function $d(\mathbf{s},\mathbf{x}_i)$ has no local maxima when restricted to an arbitrary plane $\mathbf{s}.\mathbf{n} = \lambda$, and thus if $R(\mathbf{s})$ was maximized within the plane, it must correspond to a transition between two Voronoi cells, so $d(\mathbf{s},\mathbf{x}_i) = d(\mathbf{s},\mathbf{x}_j) = d(\mathbf{s},\mathbf{p})$ for some *i* and *j*. Since **s** is equidistant from **p** and two atoms, it implies that **s** is on an edge of $V_\mathbf{p}$, violating the original assumption. Thus, it is not possible for a local maximum of $R(\mathbf{s})$ to exist on the interior of a face of $V_\mathbf{p}$.

Directly analogous steps can be taken to show that a local maximum cannot exist on an edge of $V_\mathbf{p}$: the function $d(\mathbf{s},\mathbf{x}_i)$ has no local maxima when restricted to a line, and if $R(\mathbf{s})$ is maximized on a line, it is equidistant to three atoms plus **p**, meaning that it must lie on a vertex of $V_\mathbf{p}$. The PSD algorithm therefore searches over all Voronoi network vertices and all ghost cell vertices in order to determine the maximum value of $R(\mathbf{s})$.

To take into account the atomic radii, the radical Voronoi tessellation is employed, whereby in the equations above, the distance $d(\mathbf{x},\mathbf{x}_i)$ is replaced with $d(\mathbf{x},\mathbf{x}_i)^2 - R_i^2$, where R_i is the radius of atom *i*. The sampled point **p** is, for the purposes of this calculation, defined to have radius equivalent to the probe radius. The square weighting in distances ensures that the Voronoi cells of the network and the ghost cells are irregular polyhedra with planar faces.

The case where **s** is positioned at a node of the ghost cell is illustrated in Fig. 2. Fig. 2A shows the initial radical Voronoi tessellation of the void space, and Fig. 2B shows the corresponding spheres that are centered on the nodes. The largest sphere that could possibly encapsulate the sample point, as seen in Fig. 1, is not centered on a node, and must therefore be calculated separately. Fig. 2C shows the construction of a Voronoi ghost cell and Fig. 2D shows the placement of spheres defined by nodes of the ghost cell. The largest of the latter encompasses the sample point and is displayed in purple.

The PSD calculations sample a specified number of random points and generate a histogram presenting the number of points assigned to particular pore diameters. We use 100,000 samples per

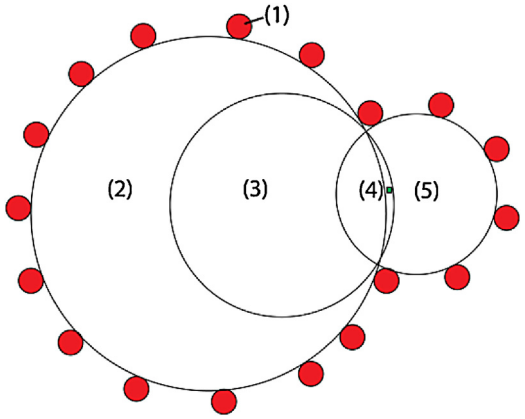

Fig. 1. Sample point where Voronoi node sphere does not represent the largest possible sphere that encapsulates sample point: (1) framework atom, (2) large node, which does not contain sample point, (3) largest possible sphere that encapsulates sample point, (4) sample point, inside small Voronoi node sphere and outside large Voronoi node sphere, and (5) small Voronoi node sphere.

structure and a histogram bin size of 0.5 Å. Our implementation of the algorithm provides three versions of the PSD histogram: the bin count, the cumulative distribution, and the derivative of the cumulative distribution. The cumulative distribution describes the probability that **p** can be found either at or below the stored diameter value, and is normalized by the maximum number of accessible samples. The derivative distribution describes the change in the cumulative distribution with respect to pore size.

2.2.2. Stochastic ray-tracing approaches

The stochastic ray tracing algorithm [42] is an alternative approach, complementary to the PSD, to generate a histogram representation of a material. In this approach, the void space is probed by randomly sampled rays within the unit cell of a material. For each ray, the distance that it travels until it hits the probe accessible surface, defined as the periphery of space accessible to the probe center, is recorded. The Monte Carlo algorithm samples a specified number of rays that originate at random points in accessible space, and creates a histogram of the resultant ray lengths.

In this algorithm, the starting point and direction of a ray are selected at random. The algorithm then proceeds to identify intersection points of the ray with the probe AS of the material, enabling lengths of rays within the pore to be calculated. We consider two versions of this approach.

In the *constrained* ray-tracing approach, sample points are only selected from within the probe accessible volume, and rays are terminated upon their first collision with the AS. The constrained approach therefore only provides information on probe-accessible regions within a material. In the *unconstrained* ray-tracing (or simply, ray-tracing) approach, any point sampled within the unit cell can give rise to a sampling ray. Additionally, the ray is permitted to continue through the AS, and may explore many voids on its path until it reaches the maximum length threshold as defined in the algorithm. Lengths of each ray fragment going through each void space segment are recorded and contribute to the overall histogram. Hence, the unconstrained approach includes some representation of a materials density, which may be desired in some applications.

All of these calculations make use of periodic boundary conditions. The above procedure is repeated until the desired number of sampled rays is reached. We sample 100,000 rays per material, with an allowed maximum ray-length of 100 Å. The ray histograms were prepared using 1000 bins of 0.1 Å size.

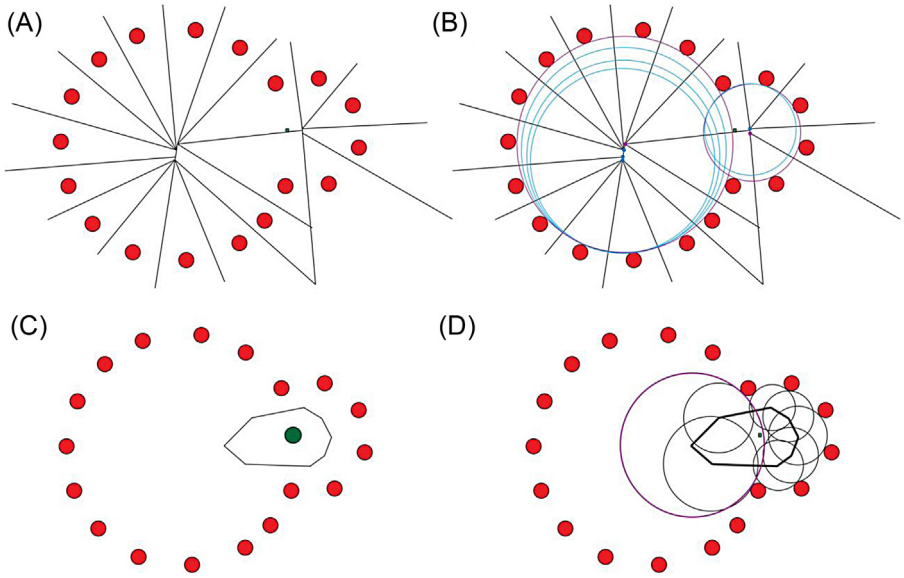

Fig. 2. (A) Voronoi tessellation of pore network in Fig. 1. Red circles represent framework atoms; green dot represents sample point; black lines represent perpendicular bisectors that form Voronoi cell edges. (B) Addition of node spheres. The blue and purple dots represent Voronoi nodes; the blue and purple circles represent the corresponding Voronoi node spheres. The purple circles signify the largest spheres that can fill this area. Note that the sample point falls within the smaller node. (C) Construction of new Voronoi "ghost" cell, using the radical Voronoi tessellation method. The green circle represents a pseudo-sphere with a radius equivalent to the probe radius, centered on the sample point. The new cell edges are the perpendicular bisectors between the pseudo-sphere and the surrounding atoms. (D) Ghost cell node spheres. The purple circle denotes the largest sphere out of these node spheres; note that it is larger than the original node sphere that encapsulates the sample point. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

2.3. Similarity searches and criteria

There is a great deal of literature [45] regarding the quantification of (dis)similarity between chemical structures, particularly small drug-like molecules, using both binary [46] and continuous [47] structural descriptors. In this work we have used two measures: similarity based on Euclidean distance [45b], and a modified version of the Tanimoto coefficient [48], MTU, which we developed previously to compare Voronoi hologram representations of void space [34].

2.3.1. Euclidean similarity

Euclidean distance is a common method for measuring the difference between two histograms:

$$dist_{Euclid} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (1)$$

where n is the histogram length, and A_i and B_i are the values in bin i of histograms A and B, respectively. The range for Euclidean distance is $[0,\infty)$; as the distance increases, histograms can be considered to be increasingly dissimilar. However, for a set of Euclidean distances, we can obtain normalized (range [0,1]) distances by dividing each value by the absolute range (*i.e.* $dist_{max} - dist_{min}$). A set of Euclidean distances can hence be converted into a set of normalized similarities:

$$E_{norm} = 1 - \left(\frac{dist_{Euclid}}{dist_{\max} - dist_{\min}}\right) \quad (2)$$

2.3.2. Modified Tanimoto similarity

The binary version of the Tanimoto similarity coefficient between two histogram representations is given by

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

where a and b represent the number of nonzero bins in A and B, respectively, and c represents the number of coincident nonzero bins. The Tanimoto coefficient can be also defined for continuous data with A_i and B_i defined previously:

$$Tan_{cont} = \frac{\sum_{i=1}^{n} A_i B_i}{\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n} B_i^2 - \sum_{i=1}^{n} A_i B_i}. \quad (4)$$

The sparseness of data in histograms, which may contain many empty bins, can lead to very low Tanimoto similarity values. Therefore, to counter this effect, we also consider coincident empty histogram bins and quantify common absence. The Tanimoto common absence coefficient is strictly binary:

$$Tan_{abs} = \frac{n + c - a - b}{n - c}. \quad (5)$$

The modified Tanimoto coefficient (MTU) [34] is obtained by combining these similarities:

$$MTU_{cont} = \frac{Tan_{cont} + Tan_{abs}}{2}. \quad (6)$$

2.4. Structural datasets

We demonstrate here the application of these void space representations in pore landscape comparisons using a dataset of 148 zeolites from the International Zeolite Association (IZA) database. This set was obtained by selecting IZA structures with free sphere diameter larger than 1.625 Å, corresponding to a CH_4 probe. All calculations of descriptors were carried out using our Zeo++ code, with 100,000 samples in all Monte Carlo algorithms. When discussing the PSD and stochastic ray approaches (Section 3.1), we used a small probe radius of 0.5 Å, as such a small probe maps details of the pore landscape. However, when discussing similarity measures (Sections 3.2 and 3.3), we used a probe radius of 1.625 Å following Ref. [34]. The radii of oxygen and silica atoms were set to 1.35 Å. In our survey of various representations and similarity measures, we included Euclidean and MTU similarity for PSD histograms, ray trace histograms, and Voronoi holograms. PSD histograms were truncated to 16 Å, as pore diameters seldom exceeded this value

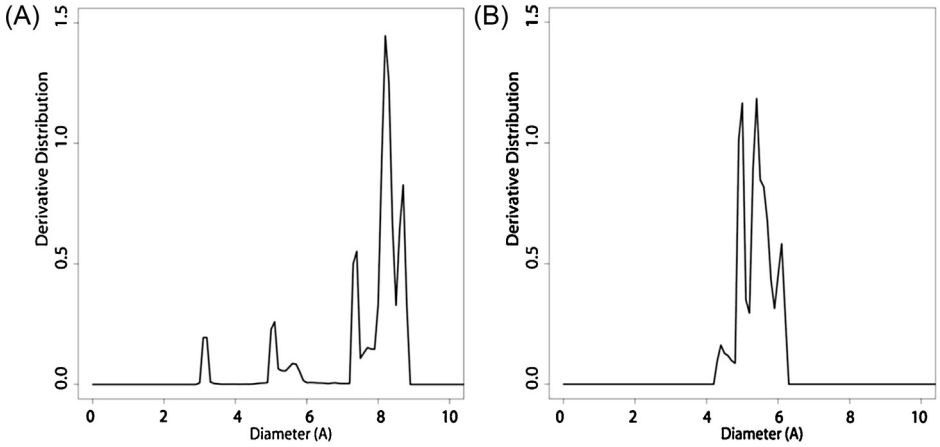

Fig. 3. PSD histograms for (A) AFT and (B) SIV. Both PSD histograms were generated with a 0.5 Å probe radii.

in our sets. For completeness we include Euclidean similarity comparisons based on one-dimensional structure representations: D_i, D_f, accessible surface area (ASA), and accessible volume (AV).

3. Results

3.1. Pore landscape characterization

We present analysis of two example zeolites, AFT and SIV, to demonstrate the capabilities and characteristics of PSD and ray tracing techniques. AFT has a non-orthogonal unit cell (13.691 Å × 13.691 Å × 29.449 Å), with a γ angle of 120°, and a three-dimensional 8-MR channel structure with larger cages between channels. SIV has an orthogonal unit cell (9.8768 Å × 14.0754 Å × 28.1314 Å) and intersecting straight and sinusoidal 8-MR channels with cages formed at their intersections. Both materials exhibit a 3-dimensional channel system accessible to a 1.625 Å radius probe. We also comment on the computational time required to generate descriptors for the 148 structure dataset, utilizing one CPU core on a standard Intel i7 desktop machine. We note that computational time varies greatly among structures, depending on factors such as channel shape and size, and unit cell parameters.

3.1.1. PSD analysis

The number of data points in a PSD histogram is equal to the number of random sample points that are within probe-accessible space. 67 min were required to generate PSD histograms for all 148 structures, with 100,000 sample points for each structure, and a 0.5 Å probe radius.

In the following examples, we generated plots of the derivative distribution (Fig. 3), which essentially provide a normalized representation of the bin count histogram. The relative heights of the peaks are equivalent to the magnitude of change in the cumulative distribution column; bin counts of 0 will not change the cumulative distribution and will therefore have a derivative distribution value of 0, while nonzero bins will have derivative distribution values that are relative to the fraction of bin count over total number of accessible points.

The PSD for AFT required 41 s to compute, and contained 43,708 accessible data points. The plot of the AFT derivative distribution (Fig. 3A) shows that AFT has five peaks at 3 Å, 5 Å, 7.3 Å, 8.2 Å, and 8.7 Å (and a small one at 5.5 Å), which correspond to the distinctive blocks of color in the AFT scatterplot (Fig. 4A). The peaks at 3 Å and 5 Å (blue and green points in Fig. 4A, respectively) are only half the height of the 7.3 Å peak (yellow points), and the 8.2 Å and 8.7 Å peaks correspond to the orange and red regions that make up the majority of the probe-accessible space.

The PSD for SIV required 28 s and contained 39,309 data points. The plot of the SIV derivative distribution (Fig. 3B) shows that SIV has four peaks at 4.3 Å, 4.8 Å, 5.4 Å, and 6.1 Å. These peaks are much less spread apart than those observed in the AFT distribution, indicating a smaller range of pore diameter sizes, with most of the points in the 5–6 Å region (green and yellow points in Fig. 4B).

3.1.2. Ray trace analysis

All ray tracing analyses presented here made use of 100,000 rays per structure and 0.5 Å probe radii. The constrained ray trace technique required 31 min to analyze all 148 structures, while the unconstrained method required more than 4 h. This significantly higher computational requirement is a consequence of tracking

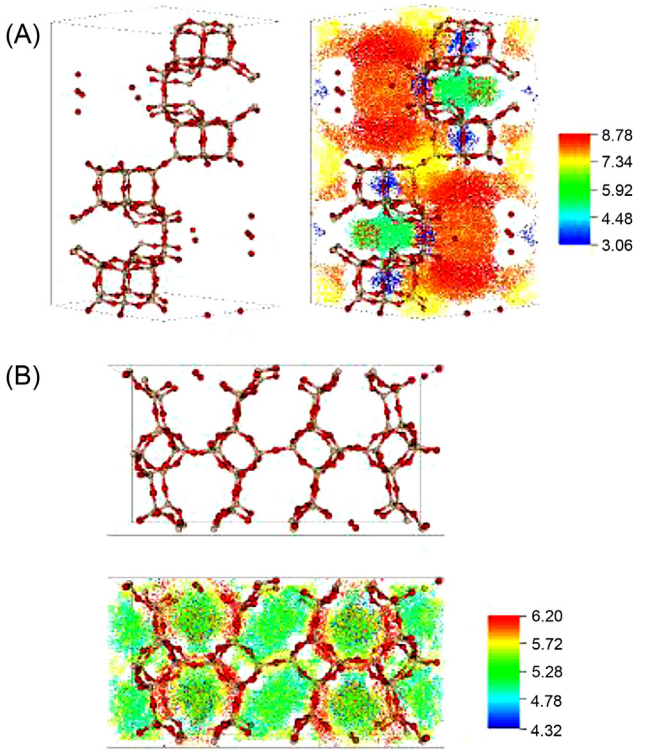

Fig. 4. PSD scatterplots for (A) AFT and (B) SIV, with corresponding pore diameters (in Å) colored according to the key.

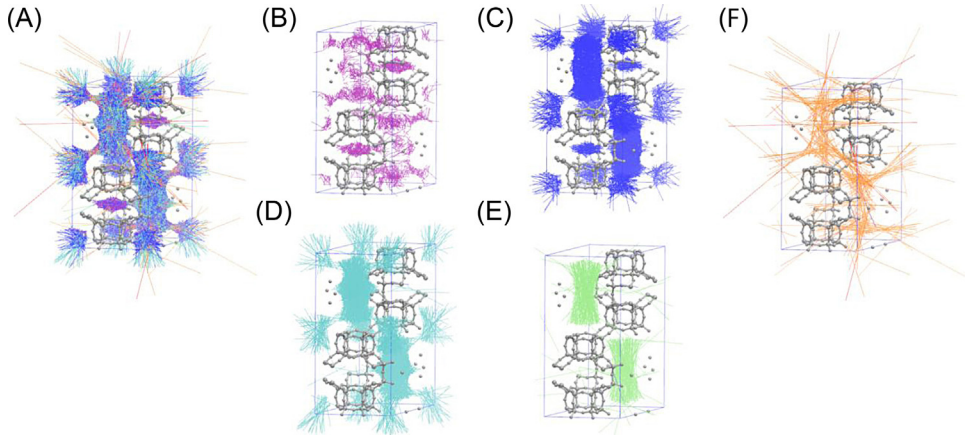

Fig. 5. Ray trace method for AFT. Lines of length (a) 0–100 Å, (b) 0–3 Å, (c) 3–6 Å, (d) 6–9 Å, (e) 9–12 Å, and (f) larger than 12 Å.

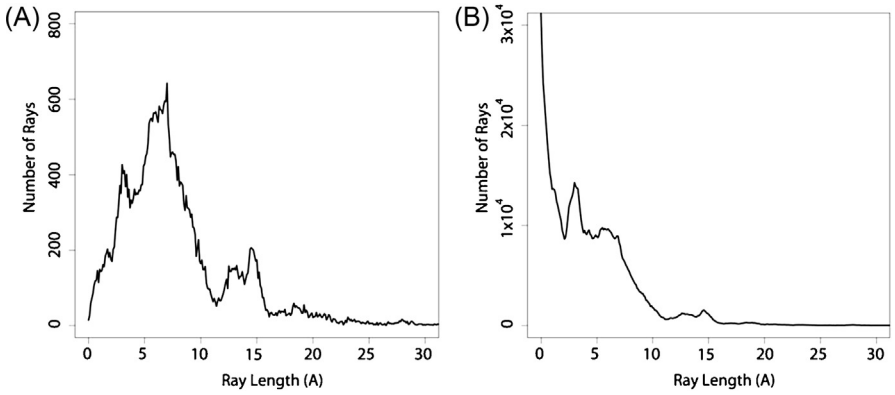

Fig. 6. AFT ray trace histograms: (A) constrained ray trace method and (B) unconstrained ray trace method.

each ray until its maximum length set up by the user (100 Å) instead of stopping the algorithm after the closest surface is reached. Figs. 5 and 7 display the rays generated by Zeo++ using the constrained method, with views B through F showing rays at various length intervals.

The constrained ray trace for AFT (Fig. 6A) required 17.01 s, whereas the unconstrained ray trace (Fig. 6B) required 242 s. As the rays are oriented in different and random directions, we are able to approximate the pore curvature, as can be seen from Fig. 5C through E. The rays in the 3–6 Å range and 6–9 Å range (Fig. 5C and D), fill the pores horizontally and diagonally, and the rays in the 9–12 Å range (Fig. 5E) fill the large pores vertically.

The constrained ray trace (Fig. 8A) for SIV required 15.56 s, whereas the unconstrained ray trace (Fig. 8B) required 237 s. The ray distribution in this structure is more uniform, with a large single distinct peak in the histogram (Fig. 8A). In Fig. 7B and C, the uniformity displayed in the histogram plot is supported by fairly even dispersal of the rays in the channels, with the exception of dense clusters at the narrower pore diameters in Fig. 7B that correspond to the cluster of blue dots seen in the PSD scatterplot in Fig. 4B.

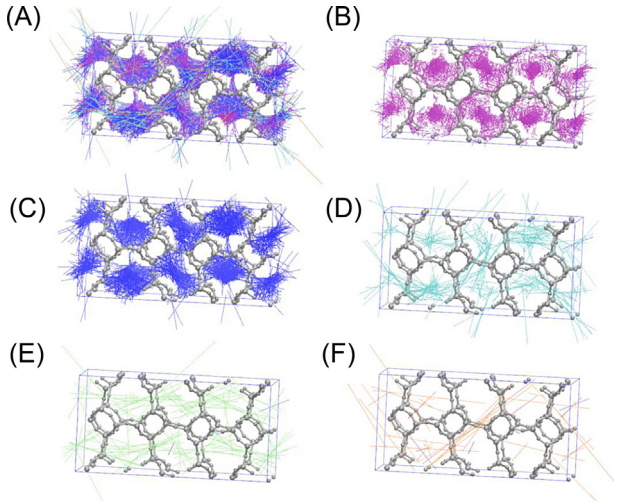

Fig. 7. Ray trace method for SIV. Lines of length (a) 0–100 Å, (b) 0–3 Å, (c) 3–6 Å, (d) 6–9 Å, (e) 9–12 Å, and (f) larger than 12 Å.

3.2. Comparing pore landscapes

PSD and ray-tracing histograms can be used to automatically compare pore landscapes of materials. They capture distinct but complementary characteristics of pores, and can therefore be used for different applications. In this section, we comment on the differences between structure comparisons using these two techniques. By means of example, we compare zeolites AFX and CHA (Fig. 9), which are similar in their constrained ray tracing histograms, and AWW and CFI (Fig. 10), which are similar in their PSD histograms. Pore landscapes, PSD, and constrained ray trace histograms for these structures are presented in Figs. 9 and 10 as well. They were calculated using 100,000 samples and a probe radius of 1.625 Å.

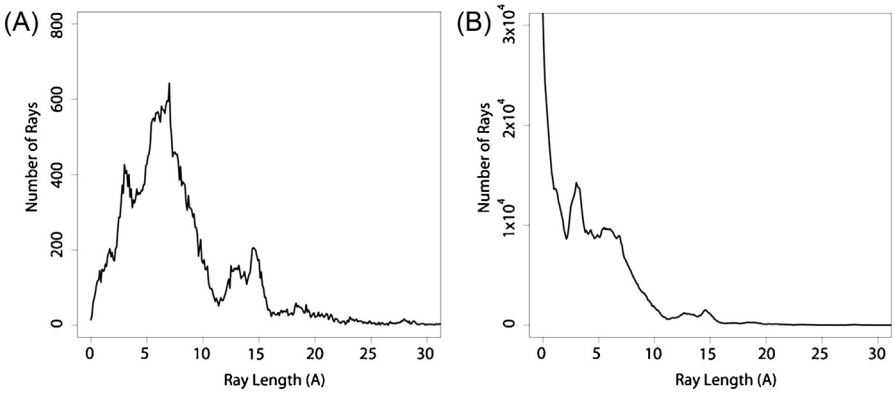

Fig. 8. SIV ray trace histograms: (A) constrained ray trace method and (B) unconstrained ray trace method.

The Euclidean similarity between constrained ray tracing histograms for AFX and CHA is 0.9195, while the corresponding MTU similarity is 0.9634. These materials appear almost identical by inspection of their ray histograms and pore landscapes (Fig. 9A, B, D and E); however, they exhibit small differences in their pore diameter. These small deviations are captured in PSD histograms (Fig. 9C and F): CHA exhibits a single peak at around 8 Å, whereas AFX exhibits two smaller peaks. This subtle difference in PSD is reflected in smaller PSD similarity values: the Euclidean similarity between AFX and CHA is 0.7246, whereas the corresponding MTU similarity is 0.5436. We observe, then, that the PSD is much more sensitive to subtle differences in pore size than the ray tracing method. Accordingly, comparing materials by their PSD histograms is a superior choice if pore *size* variations are important. However, it is important to note that the pore size distribution derived pore size is a product of a spherical representation of the pore such that the ray tracing method may be more informative if pores do not accommodate the spherical probe shape and if the overall pore *shape* comparison is considered more important. This example also highlights the difference between the Euclidean and MTU similarity measures: MTU exhibits smaller similarity values, and therefore detects subtler structural differences.

CFI and AWW, as seen in Fig. 10, are composed of approximately spherical cages connected by narrow channels. The cages exhibit nearly identical size, as shown by the PSD histograms (Fig. 10C and F): the PSD Euclidean similarity value is 0.9486, and the

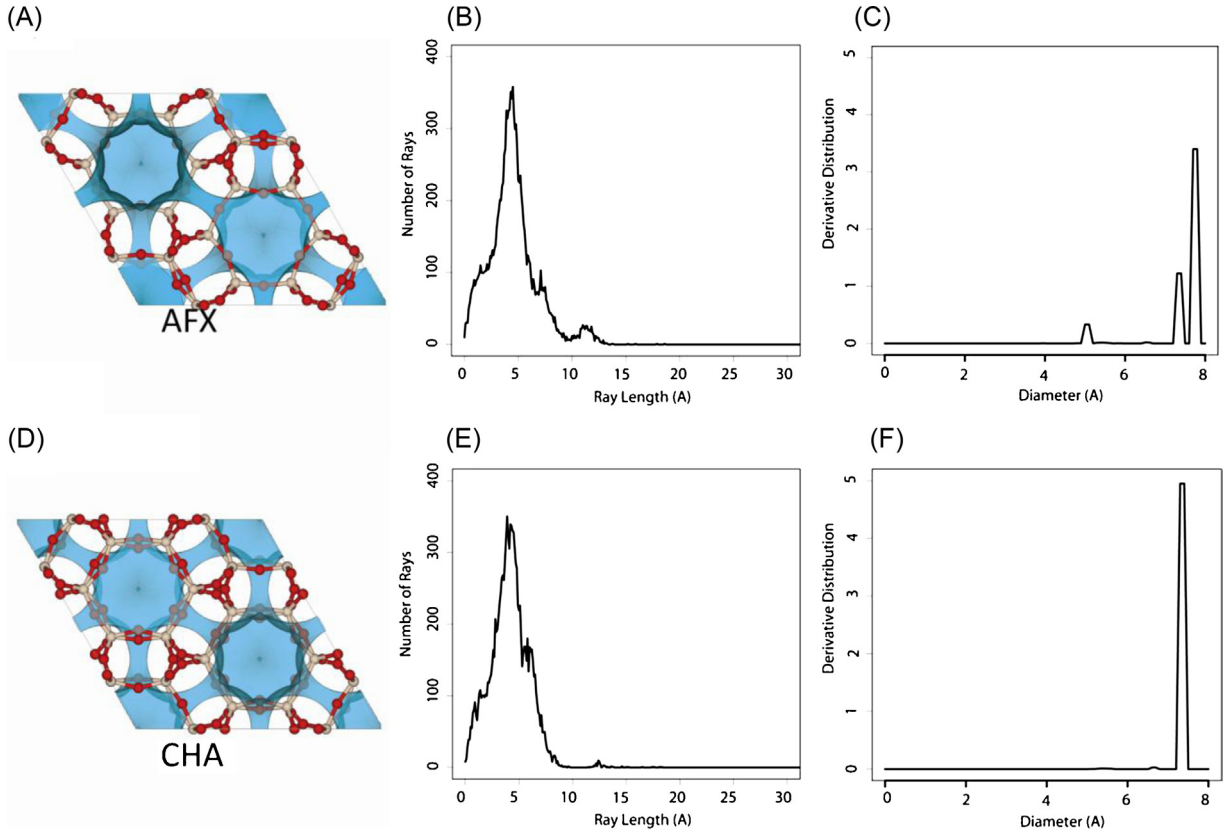

Fig. 9. Comparison of pore landscapes: (A) AFX pore landscape with (B) constrained ray trace histogram and (C) PSD histogram, and (D) CHA pore landscape with (E) constrained ray trace histogram and (F) PSD histogram. All figures generated with respect to 1.625 Å radii probes.

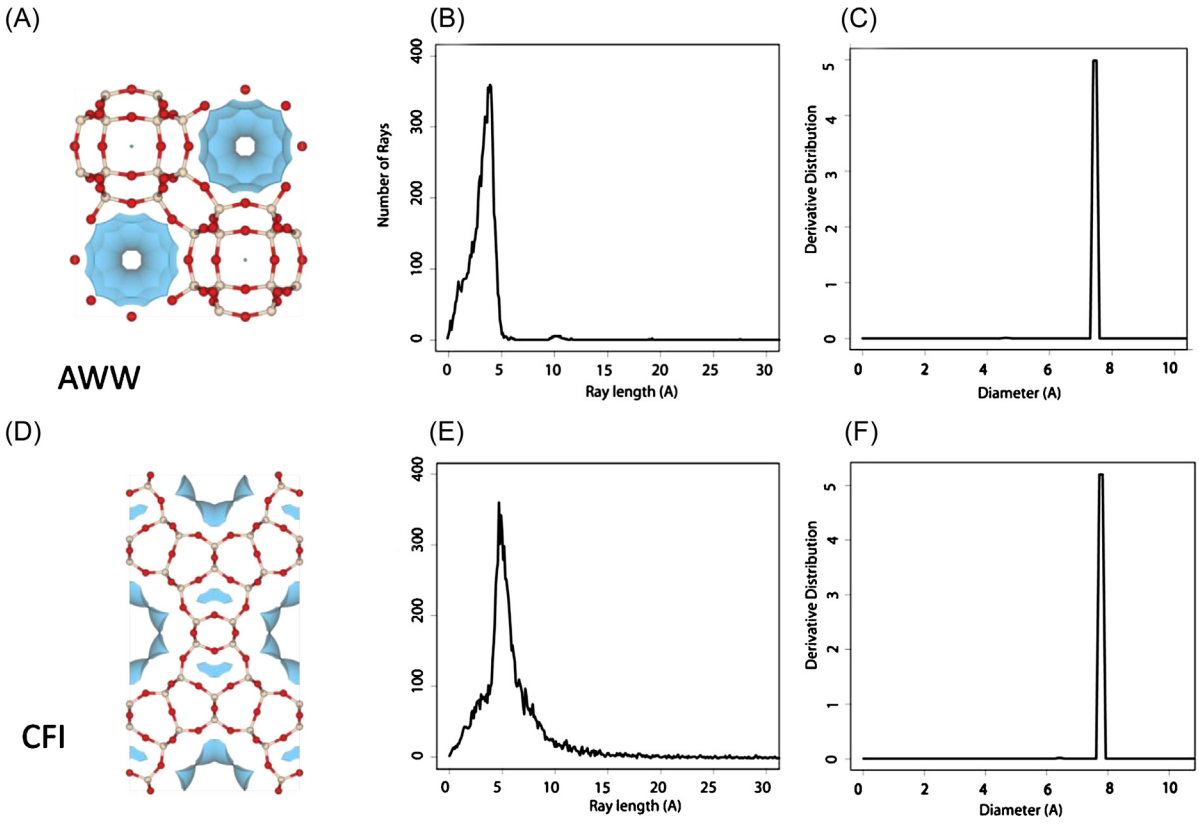

Fig. 10. Comparison of pore landscapes: (A) AWW pore landscape with (B) constrained ray trace histogram and (C) PSD histogram, and (D) CFI pore landscape with (E) constrained ray trace histogram and (F) PSD histogram. All figures generated with respect to 1.625 Å radii probes.

corresponding MTU value is 0.9400. Despite their similar pore size, these two structures exhibit very different pore surface *textures*, with the pore of AWW being smoother than CFI. This difference in pore surface is easily observed in the constrained ray tracing histograms (Fig. 10B and E), and their calculated similarities of 0.6869 and 0.5925 for Euclidean and MTU, respectively. This comparison serves as a contrast to the previous example; here, we observe greater sensitivity to small structural differences and therefore lower similarities with the ray tracing approach. We illustrate by these comparisons that each descriptor captures different structural information, and that the choice of descriptor chosen to classify and compare materials must be appropriate to the requirements of a specific application.

3.3. Similarity measures

In the previous section, we demonstrated how two representations of microporous solids can capture different structural features. With a broad choice of descriptors and representation, together with a range of similarity measures, it is essential to know which approaches can provide quantitatively new ways to compare materials. Therefore it is important to consider how the similarities

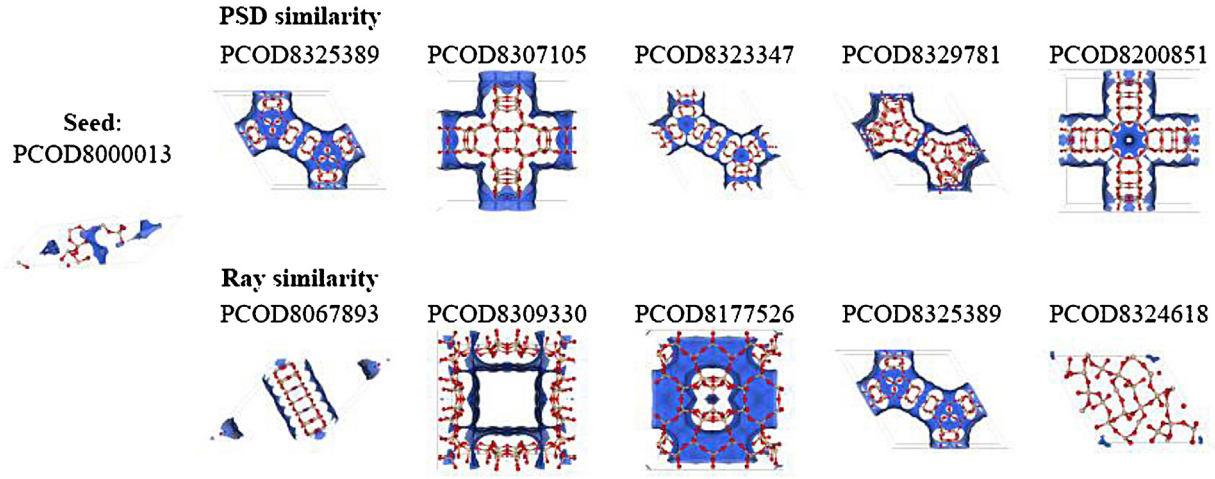

Fig. 11. Visualization of first 5 structures from heuristic MMDP diversity selection, for PSD (top row) and rays (bottom row).

Table 1
Correlation matrix: Pearson's similarity coefficients for 1D descriptors (included sphere (D_i), free sphere (D_f), accessible surface area (ASA) and accessible volume (AV)) using Euclidean distance (E); and holograms (H), ray trace methods (unconstrained (URT) and constrained (CRT)), and PSD coefficients using both Euclidean distance (E) and modified unweighted Tanimoto (MTU) coefficients.

D_i(E)	0.3189	0.3064	0.6956	0.2781	0.1790	0.7106	0.7707	0.5740	0.4736	0.4713	0.4834
0.3189	**D_f(E)**	0.0098	0.2317	0.2021	0.0269	0.5307	0.3657	0.6681	0.3665	0.2275	0.2568
0.3064	0.0098	**ASA(E)**	0.6658	0.0876	0.0087	0.2729	0.4328	0.2346	0.1974	0.1594	0.3596
0.6956	0.2317	0.6658	**AV(E)**	0.1601	0.0872	0.6268	0.8299	0.4896	0.3472	0.3065	0.6361
0.2781	0.2021	0.0876	0.1601	**H(MTU)**	0.1775	0.5104	0.4099	0.4741	0.4999	0.4885	0.2002
0.1790	0.0269	0.0087	0.0872	0.1775	**H(E)**	0.0517	0.1128	0.0699	0.0942	0.1361	0.0977
0.7106	0.5307	0.2729	0.6268	0.5104	0.0517	**CRT(MTU)**	0.8441	0.8586	0.7231	0.5093	0.4091
0.7707	0.3657	0.4328	0.8299	0.4099	0.1128	0.8441	**CRT(E)**	0.6909	0.5689	0.4892	0.6865
0.5740	0.6681	0.2346	0.4896	0.4741	0.0699	0.8586	0.6909	**URT(MTU)**	0.8142	0.4120	0.3725
0.6268	0.3665	0.1974	0.3472	0.4999	0.0942	0.7231	0.5689	0.8142	**URT(E)**	0.3896	0.2017
0.4713	0.2275	0.1594	0.3065	0.4885	0.1361	0.5093	0.4892	0.4120	0.3896	**PSD(MTU)**	0.6013
0.4834	0.2568	0.3596	0.6361	0.2002	0.0977	0.4091	0.6865	0.3725	0.2017	0.6013	**PSD(E)**

Table 2
First 5 structures from heuristic MMDP (diversity selection), with structure PCOD8000013 as seed.

		Structure 1	Structure 2	Structure 3	Structure 4	Structure 5
PSD MTU similarity	Structure	PCOD8325389	PCOD8307105	PCOD8323347	PCOD8329781	PCOD8200851
	Similarity	0.2195, with PCOD8000013	0.2966, with PCOD8000013	0.3051, with PCOD8000013	0.3305, with PCOD8000013	0.3362, with PCOD8000013
Ray MTU similarity	Structure	PCOD8067893	PCOD8309330	PCOD8177526	PCOD8325389	PCOD8324618
	Similarity	0.0224, with PCOD8000013	0.1243, with PDOC8067893	0.3023, with PCOD8309330	0.3231, with PCOD8309330	0.3806, with PCOD8000013

obtained with each descriptor are correlated. In this work, we have considered Euclidean similarity and modified Tanimoto similarity coefficients. The choice of coefficient determines the similarity obtained when comparing descriptors, indicating that one must select an appropriate similarity coefficient as well as an appropriate descriptor. It is worth noting that overall, there were varying degrees of correlation between the one-dimensional and the multi-dimensional variables. Descriptors such as D_i or D_f are limited to very specific comparisons, and a low correlation value between D_i and PSD makes sense when one considers that similarity between sphere diameters fails to describe the rest of the unit cell. The degree of correlation between similarity measures depends upon the distribution of similarity values in a dataset, and choosing the appropriate method for calculating similarity requires an understanding of what the similarity coefficients are quantifying.

Similarity based on normalized Euclidean distance is sensitive to large outliers in a dataset, due to the normalization requirement. For example, we observe that the distribution of Euclidean similarities between Voronoi holograms was heavily skewed toward 1, with a small percentage at or near 0. This is the result of a large proportion of Euclidean distances being small: in the distribution of raw distance values, most were clustered in the 0–1000 range, with a smaller number in the 1000–2000 range and a tiny fraction in the 5000 range. However, because the distance values are normalized by the maximum distance, all values were divided by 5184.43, so the majority of normalized distances were in the 0.1–0.2 range, yielding a majority of similarity coefficients in the 0.8–0.9 range for holograms utilizing the Euclidean distance method. Thus, a few outliers (most likely the distances between the smallest and largest structures) affected the similarity coefficients of the entire set.

As a general observation, it is important to note that since normalized Euclidean similarity between two materials is affected by the range of Euclidean distances observed in the set, it will be altered as new materials exhibiting greater distances are added to the dataset. A similarity coefficient, such as MTU, does not exhibit this behavior; the MTU similarity between two descriptors is a constant. However, while MTU does not have the same sensitivity to outliers and new data points as Euclidean distance, the distribution of the MTU coefficient, due to the binary nature of the Tan_{abs} component, could be skewed in the positive or negative direction by common absence; indeed, this skew was observed for the PSD similarities using MTU, whereas Euclidean similarities provide a more normal distribution.

The differing Euclidean distance and MTU similarity coefficients indicate that the two methods are not interchangeable. Using the Pearson correlation coefficient [49], we have examined the correlation of descriptor similarities, obtained for each similarity measure (Table 1, also Supporting Data), and have determined that not only is the correlation between Euclidean distance and MTU less than 1 for all descriptors, but that the amount of correlation is also variable. This raises a fundamental question: what is the definition of similarity in the context of these descriptors? Euclidean distance does not explicitly detect common presence/absence, since distance values do not specify whether bins are full or empty; a distance value of 1000 could be the difference between 3000 and 2000 or the difference between 1000 and 0. MTU addresses this situation, but in restricting similarity to common features, it might miss features that are close in size, but not exactly alike.

The contrast between these two methods was particularly evident in the case of holograms, where the Pearson's correlation between MTU and Euclidean similarities is only 0.1775. Overall, the Euclidean hologram similarities had the lowest correlation values with respect to the other similarity coefficients, while the ray trace similarities had some of the highest correlation values with respect to the other similarity coefficients. In some circumstances, the Tanimoto coefficient of a descriptor exhibited a higher correlation with the one-dimensional descriptors, while in other cases the Euclidean distance coefficient exhibited a higher correlation value. Therefore, we cannot definitively state that either similarity measure can be applied in all circumstances; choosing the correct method requires insight into the nature of the data (possibility of outliers, distribution of raw data, number of empty bins, etc.).

3.4. Diversity selection

Beginning with the structure that was alphanumerically first on the list of hypothetical zeolites, PCOD8000013, we utilized the heuristic MMDP algorithm to obtain a list of the 1,000 most diverse structures in the PCOD dataset, with computation time of roughly 4.5 h per descriptor.

In the preceding sections, we discussed how PSD and ray tracing methods select for different types of structural characteristics of materials. For the IZA dataset, we had calculated a Pearson's correlation coefficient of 0.5093 between the two descriptors' MTU similarity values, so it is logical to assume that we would see differences in diversity selection as well. Out of the 1000 structures per diversity selection subset, there were only 88 duplicates between the two lists (including the seed structure). In the PSD diversity subset, similarity values range from 0.2195 to 0.8743 (a range of 0.6548); in the ray diversity subset, similarity values range from 0.0224 to 0.6751 (a range of 0.6527). Table 2 and Fig. 11 respectively display the MTU similarity values and structures for the first five selections in the PSD and ray MMDP subsets. We observe that, from a purely visual standpoint, the diverse structures identified using rays appear to be more dissimilar to one another than those identified using PSD. However, we emphasize that despite multiple structures exhibiting a similar overall pore layout, the structures in the PSD diverse set exhibit pores and connecting channel systems of differing sizes.

4. Conclusions

We have presented pore size distribution and stochastic ray tracing techniques for characterizing void space within a porous material and have demonstrated that these techniques provide complementary information about the pore structure. PSD describes pore landscapes with lower and upper bounds of pore diameters and their relative proportions, and accordingly is highly sensitive to small changes in pore diameter; however, it does not reflect subtle changes in features such as the surface texture of a pore. By comparison, stochastic ray trace histograms describe pore landscapes in terms of overall shape, and reflect a subjective visual assessment of a material's pore shape; ray tracing captures subtle deviations in pore texture, but is not as sensitive to deviation in pore diameter and features can often be obscured by overlap in histograms. These new tools enable a materials researcher to assess pore structure from multiple perspectives, without the need for visualization of individual structures, in an automated and high-throughput manner. Moreover, these sophisticated descriptors can be utilized for material analysis needs that are not adequately met by conventional one-dimensional, numerical descriptors such as restricting pore diameter (D_f).

Additionally, we have discussed the performance of these new descriptors under two distinct similarity measures, and in comparison to existing pore descriptors. These comparisons have illustrated that the choice of pore descriptor and similarity measure has a large influence on the perspective of material similarity obtained with each descriptor exhibiting a particular emphasis and sensitivity to certain aspects of structural similarity. The addition of these new

tools to the array of techniques available for porous material analysis will enable researchers to make an informed choice on how to characterize and screen material databases, and to efficiently and automatically discover materials exhibiting the particular pore characteristics of their interest.

Acknowledgements

RLM and MH were supported by the US Department of Energy under Contract No. DE-AC02-05CH11231. In addition, it was supported jointly by DOE Office of Basic Energy Sciences through project #CSNEW918 entitled "Knowledge guided screening tools for identification of porous materials for CO_2 separations", and as part of the Center for Gas Separations Relevant to Clean Energy Technologies, an Energy Frontier Research Center funded by the US 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 US Department of Energy under Contract No. DEAC02-05CH11231.

AJ and EI acknowledge the financial support of the Chevron Energy Technology Company.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.05.007.

References

[1] T.F. Degnan Jr., The implications of the fundamentals of shape selectivity for the development of catalysts for the petroleum and petrochemical industries, J. Catal. 216 (2003) 32–46.
[2] M. Dekker, Handbook of Zeolite Science and Technology, New York, 2004.
[3] B. Smit, T.L.M. Maesen, Towards a molecular understanding of shape selectivity, Nature 451 (2008) 671–677.
[4] B. Smit, T.L.M. Maesen, Molecular simulations of zeolites: adsorption, diffusion, and shape selectivity, Chem. Rev. 108 (2008) 4125–4184.
[5] R. Krishna, J.M. van Baten, Using molecular simulations for screening of zeolites for separation of CO_2/CH_4 mixtures, Chem. Eng. J. 133 (2007) 121–131.
[6] A.R. Millward, O.M. Yaghi, Metal–organic frameworks with exceptionally high capacity for storage of carbon dioxide at room temperature, J. Am. Chem. Soc. 127 (2005) 17998–17999.
[7] K.S. Walton, A.R. Millward, D. Dubbeldam, H. Frost, J.J. Low, O.M. Yaghi, R.Q. Snurr, Understanding inflections and steps in carbon dioxide adsorption isotherms in metal–organic frameworks, J. Am. Chem. Soc. 130 (2008) 406–407.
[8] R. Banerjee, A. Phan, B. Wang, C. Knobler, H. Furukawa, M. O'Keeffe, O.M. Yaghi, High-throughput synthesis of zeolitic imidazolate frameworks and application to CO_2 capture, Science 319 (2008) 939–994.
[9] K. Sumida, M.R. Hill, S. Horike, A. Dailly, J.R. Long, Synthesis and hydrogen storage properties of $Be1_2(OH)_{12}$(1,3,5-benzenetribenzoate)$_4$, J. Am. Chem. Soc. 131 (2009) 15120–15121.
[10] H.J. Choi, M. Dinca, J.R. Long, Broadly hysteretic H_2 adsorption in the microporous metal–organic framework Co(1,4-benzenedipyrazolate), J. Am. Chem. Soc. 130 (2008) 7450–7848.
[11] D.M. D'Alessandro, B. Smit, J.R. Long, Carbon dioxide capture: prospects for new materials, Angew. Chem. Int. Ed. 49 (2010) 6058–6082.
[12] (a) S. Yang, M. Lach-hab, I.I. Vaisman, E. Blaisten-Barojas, X. Li, V.L. Karen, Framework-type determination for zeolite structures in the inorganic crystal structure database, J. Phys. Chem. Ref. Data 39 (2010) 033102–33145;
(b) S. Yang, M. Lach-hab, I.I. Vaisman, E. Estela Blaisten-Barojas, Identifying zeolite frameworks with a machine learning approach, J. Phys. Chem. C 113 (2009) 21721–21725.
[13] M.D. Foster, M.M.J. Treacy, http://www.hypotheticalzeolites.net (accessed 13.11.2009).
[14] D.J. Earl, M.W. Deem, Toward a database of hypothetical zeolite structures, Ind. Eng. Chem. 45 (2006) 5449–5454.
[15] M.W. Deem, R. Pophale, P.A. Cheeseman, D.J. Earl, Computational discovery of new zeolite-like materials, J. Phys. Chem. C 113 (2009) 21353–21360.
[16] R. Pophale, P.A. Cheeseman, M.W. Deem, A database of new zeolite-like materials, Phys. Chem. Chem. Phys. 13 (2011) 12407–12412.
[17] N.W. Ockwig, O. Delgado-Friedrichs, M. O'Keeffe, O.M. Yaghi, Reticular chemistry: occurrence and taxonomy of nets and grammar for the design of frameworks, Acc. Chem. Res. 38 (2005) 176–182.
[18] C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal–organic frameworks, Nat. Chem. 4 (2012) 83–89.
[19] B. Smit, R. Krishna, Molecular simulations in zeolitic process design, Chem. Eng. Sci. 58 (2003) 557–568.
[20] L-C. Lin, A. Berger, R.L. Martin, J. Kim, J. Swisher, K. Jariwala, C.H. Rycroft, A. Bhown, M.W. Deem, M. Haranczyk, B. Smit, In silico screening of carbon-capture materials, Nat. Mater. 11 (2012) 633–641.
[21] M.D. Foster, I. Rivin, M.M.J. Treacy, O. Delgado, A geometric solution to the largest-free-sphere problem in zeolite frameworks, Microporous Mesoporous Mater. 90 (2006) 32–38.
[22] E. Haldoupis, S. Nair, D.S. Sholl, Efficient calculation of diffusion limitations in metal organic framework materials: a tool for identifying materials for kinetic separations, J. Am. Chem. Soc. 132 (2010) 7528–7539.
[23] H. Li, A. Laine, M. O'Keeffe, O.M. Yaghi, Supertetrahedral sulfide crystals with giant cavities and channels, Science 283 (1999) 1145–1147.
[24] T. Düren, F. Millange, G. Férey, K.S. Walton, R.Q. Snurr, Calculating geometric surface areas as a characterization tool for metal–organic frameworks, J. Phys. Chem. C 111 (2007) 15350–15356.
[25] T. Düren, L. Sarkisov, O.M. Yaghi, R.Q. Snurr, Design of new materials for methane storage, Langumir 20 (2004) 2683–2689.
[26] D.D. Do, L.F. Herrera, H.D. Do, A new method to determine pore size and its volume distribution of porous solids having known atomistic configuration, J. Colloid Interface Sci. 328 (2008) 110–119.
[27] V.A. Blatov, O. Delgado-Friedrichs, M. O'Keeffe, D.M. Proserpio, Three-periodic nets and tilings: natural tilings for nets, Acta Cryst. A63 (2007) 418–425.
[28] (a) E.L. First, C.A. Floudas, MOFomics computational pore characterization of metal–organic frameworks, Microporous Mesoporous Mater. 165 (2013) 32–39;
(b) E.L. First, C.E. Gounaris, J. Wei, C.A. Floudas, Computational characterization of zeolite porous networks: an automated approach, Phys. Chem. Chem. Phys. 13 (2011) 17339–17358.
[29] L. Sarkisov, A. Harrison, Computational structure characterisation tools in application to ordered and disordered porous materials, Mol. Sim. 37 (2011) 1248–1257.
[30] M. Haranczyk, J.A. Sethian, Automatic structure analysis in high-throughput characterization of porous materials, J. Chem. Theory Comput. 6 (2010) 3472–3480.
[31] R.L. Martin, Prabhat, D.D. Donofrio, J.A. Sethian, M. Haranczyk, Accelerating analysis of void space in porous materials on multicore and GPU platforms, Int. J. High Perform. Comput. Appl. 26 (2012) 347–357.
[32] T.F. Willems, C.H. Rycroft, M. Kazi, J.C. Meza, M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials, Microporous Mesoporous Mater. 149 (2012) 134–141.
[33] A. Okabe, B. Boots, K. Sugihara, S. Nok Chiu, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Wiley, 2000.
[34] R.L. Martin, B. Smit, M. Haranczyk, Addressing challenges of identifying geometrically diverse sets of crystalline porous materials, J. Chem. Inf. Model. 52 (2012) 308–318.
[35] Zeo++, available at: http://www.carboncapturematerials.org/Zeo++ (February 1, 2013).
[36] (a) Original library available at http://math.lbl.gov/voro++/ (June 1st, 2010).;
(b) The modified library discussed here is available from the authors upon request.
[37] C.H. Rycroft, Voro++: A Three-Dimensional Voronoi Cell Library in C++. Paper LBNL-1430E, Lawrence Berkeley National Laboratory, 2009 (January 23rd 2009).
[38] V.A. Blatov, Voronoi–Dirichlet polyhedra in crystal chemistry: theory and applications, Cryst. Rev. 10 (2004) 249–318.
[39] V.A. Blatov, A.P. Shevchenko, Polyhedral representation of voids in the sodium-free sublattice of the crystal structure of $NaAlSiO_4$ zeolite, Acta Cryst. A59 (2003) 34–44.
[40] M.G. Alinchenko, A.V. Anikeenko, N.N. Medvedev, V.P. Voloshin, M. Mezei, P. Jedlovszky, Computer simulation study of intermolecular voids in unsaturated phosphatidylcholine lipid bilayers, J. Phys. Chem. B 108 (2004) 19056–19067.
[41] V.A. Blatov, G.D. Ilyushin, O.A. Blatova, N.A. Anurova, A.K. Ivanov-Schits, L.N. Dem'yanets, Analysis of migration paths in fast-ion conductors with Voronoi–Dirichlet partition, Acta Cryst. B 62 (2006) 1010–1018.
[42] A. Jones, C. Ostrouchov, M. Haranczyk, E. Iglesia, From rays to structures: representation and selection of void structures in zeolites using stochastic methods, Microporous Mesoporous Mater. (2013) (submitted for publication).
[43] S.C. van der Marck, Network approach to void percolation in a pack of unequal spheres, Phys. Rev. Lett. 77 (1996) 1785–1788.
[44] E.W. Dijkstra, A note on two problems in connexion with graphs, Numerische Mathematik 1 (1959) 269–271.
[45] (a) P.H.A. Sneath, P.R. Sokal, Numerical Taxonomy, W.H. Freeman, San Francisco, 1973;
(b) P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996;
(c) D.R. Flower, On the properties of bit string-based measures of chemical similarity, J. Chem. Inf. Comput. Sci. 38 (1998) 379–386.
[46] M. Haranczyk, J. Holliday, Comparison of similarity coefficients for clustering and compound selection, J. Chem. Inf. Model. 48 (2008) 498–508.

[47] A. Al Khalifa, M. Haranczyk, J. Holliday, Comparison of non-binary similarity coefficients for similarity searching, clustering and compound selection, J. Chem. Inf. Model. 49 (2009) 1193–1201.
[48] (a) T.T. Tanimoto, IBM Internal Report, 17th November 1957, 1957; (b) P. Jaccard, Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 241–272; (c) M.A. Fligner, J.S. Verducci, P.E. Blower, Jaccard/Tanimoto similarity index for diverse selection of chemical compounds using binary strings, Technometrics 44 (2002) 110–119.
[49] K. Pearson, Mathematical contributions to the theory of evolution. III: regression, heredity, and panmixia, Philos. Trans. R. Soc. Lond., Ser. A 187 (1896) 253–318.