Correlative Analysis of Metal Organic Framework Structures through Manifold Learning of Hirshfeld Surfaces

Journal:	*Molecular Systems Design & Engineering*
Manuscript ID	ME-ART-03-2018-000014.R2
Article Type:	Paper
Date Submitted by the Author:	01-Aug-2018
Complete List of Authors:	Shen, Xiaozhou; University at Buffalo - The State University of New York Zhang, Tianmu; University at Buffalo - The State University of New York, EE Broderick, Scott; University at Buffalo - The State University of New York, Rajan, Krishna; University at Buffalo - The State University of New York

Design, System, Application

The linking of molecular building units with covalent bonds has opened the door to a data rich field of reticular materials design[1-4]. The combinations of potential chemistries and bonding arrangements permits large numbers of materials chemistries covering a broad range of crystallographic space groups of extended structures. The resulting framework structures, such as Metal-organic framework (MOF) structures, can be tuned to create well defined pore geometries, networks and sizes. In this paper, we present a machine learning / data driven approach to characterize MOF structures. Thisapproach permits one to map high dimensional correlations between diverse chemistries and crystallography so as to discover how the interplay between chemical bonding and chemical geometry govern relationships between seemingly diverse families of MOF structures. It is demonstrated that this correlative analysis approach can serve as a way to map design strategies for MOF structures for the tuning of pore size and pore network geometry and establishes a template for interrogating databases to uncover new relationships in materials chemistries.

Correlative Analysis of Metal Organic Framework Structures through Manifold Learning of Hirshfeld Surfaces

Xiaozhou Shen, Tianmu Zhang, Scott Broderick and Krishna Rajan
Department of Materials Design and Innovation
University at Buffalo- the State University of New York
Buffalo, NY 14260-1660 USA

Abstract

We demonstrate the use of non-linear manifold learning methods to map the connectivity and extent of similarity between diverse metal-organic framework (MOF) structures in terms of their surface areas by taking into account both crystallographic and electronic structure information. The fusing of geometric and chemical bonding information is accomplished by using 3-dimensional Hirshfeld surfaces of MOF structures, which encode both chemical bonding and molecular geometry information. A comparative analysis of the geometry of Hirshfeld surfaces is mapped into a low dimensional manifold through a graph network where each node corresponds to a different compound. By examining nearest neighbor connections, we discover structural and chemical correlations among MOF structures that would not have been discernible otherwise. Examples of the types of information that can be uncovered using this approach are given.

1 Introduction

In the case of MOF systems, machine learning techniques have discovered promising materials for CH_4 uptake[5, 6] and carbon green-house gas capture[7]. In those applications, the desired properties were affected by a number of factors and their synergistic effects, such as chemical formulas, bonding (electronic structures), intermolecular packing, crystallographic parameters, the nature of the building blocks, and their network connections[6-13], were explored.

In the present study, we take an approach that is different from those previous studies in two fundamental ways:

i) We integrate geometric and chemical bonding information in the characterization of MOF structures. This integrated information is treated as our training data set for a machine learning / informatics analysis. This integration is accomplished by representing MOF structures in terms of 3-dimensional Hirshfeld surfaces that encode both chemical bonding and molecular geometry information[14-15].

ii) The correlations between compounds in terms of their structural similarities and properties are mapped from a manifold learning based analysis of the Hirshfeld surfaces (Figure 1). The information is projected in terms of a graph network that permits us to quickly identify the extent of similarity between different MOF structures and the nature of the connections or lack thereof between compounds. The ability to map the connectivity between structure, bonding, and intermolecular interactions has permitted us to now show how a new type of structure map for MOF systems can be derived.

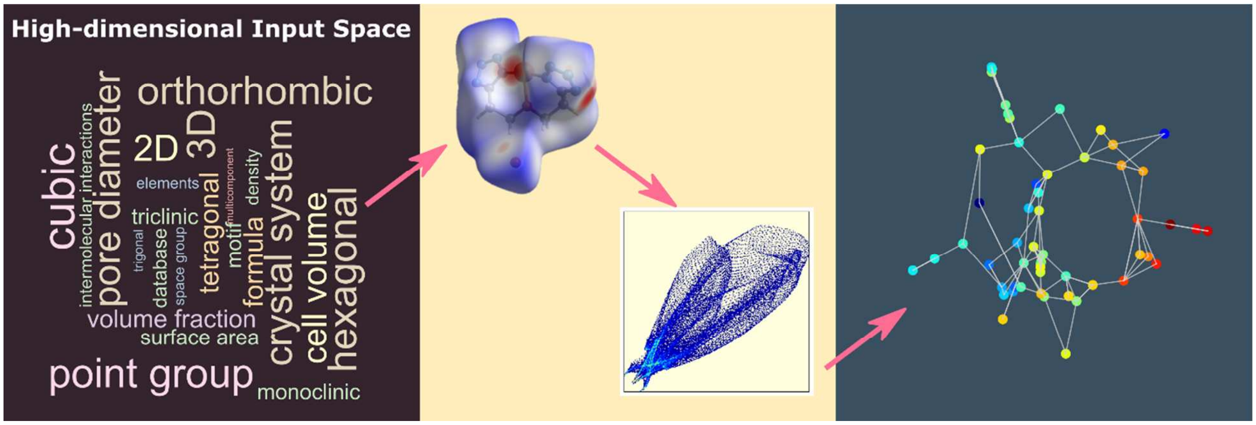

Figure 1 Schematic of the manifold learning process of MOFs structures

2 Methods

2.1 Hirshfeld surface calculations

A fundamental issue in materials design process is how to assemble molecules into molecular crystals. The idea of Hirshfeld surfaces arises from the above question: how to assemble molecules and fabrication of flexible or rigid, organic or metal-organic building blocks into multicomponent system? Hirshfeld surfaces have been widely used in studying organic and inorganic solids, including MOF structures[8-11]. It is based promolecule and procrystal models, which are rooted in quantum mechanics and simplified further by spherical-atom approximation. Promolecule and procrystal models are originated from in electron charge density analysis and X-ray diffraction analysis as the ‘independent atom model’ (IAM). In standard structural analysis of X-ray diffraction data, the scattering of a crystal is the sum of scattering by the spherically-symmetrical, ground-state atoms calculated by Hartree-Fock method. In charge density analysis, they are employed as ideal reference systems, which are made up by non-interacting atoms held fixed at the same positions as they are in the corresponding ‘real’ molecule and crystal. As the charge density of an isolated atom is spherically symmetric, the corresponding promolecular density is just the sum of the spherically-averaged atomic charge densities, each centered on the coordinates of the corresponding ‘real’ nucleus. On the other hand, procrystal density can be obtained by the same summation over the whole unit cell of a crystal.

Molecular Hirshfeld surfaces in crystal structures are constructed based on the electron distribution calculated as the sum of spherical atomic electron densities as an isosurface around molecules-in-crystals, where the density from the promolecule contribution to the procrystal density exceeds the contributions coming from all the other molecules in the periodic system (Equation 1). Hirshfeld surfaces of each molecule of a given crystal structure are unique, thus they can be used as a generic ‘finger-print’ of a particular molecule-in-crystal.

Equation 1

$$\omega(r) = \frac{\rho_{promolecule}}{\rho_{procrystal}} = \frac{\sum_{a \in molecule} \rho_a(r)}{\sum_{a \in crystal} \rho_a(r)},$$

where promolecule / procrystal: is a model of a molecule / crystal where the electron density distributions of each of its atoms have been spherically averaged and placed at their minimum energy positions.

In addition to the Hirshfeld surface itself, we can encode other geometric properties on surfaces. In particular, each point on the surface can be mapped with a set of values based on the neighboring environments. The contact distances d_e and d_i are the distances from the Hirshfeld surface to the nearest atoms outside and inside the surface, respectively. The pairs of contact distances highlight the donor-acceptor in the crystal so that they can be applied as a powerful tool for analyzing directional intermolecular interactions. The 2D histogram generated by the d_e and d_i pairs for all of the points on the Hirshfeld surface serve as a 2D fingerprint plot which provides a summary of the intermolecular contacts in the crystal.

Aside from the 2D fingerprint plots, d_i and d_e can be normalized as an indicator of intermolecular contact distance by taking the atomic van der Waal’s radii into account.

The normalized contact distance is defined as:

Equation 2

$$d_{norm} = \frac{d_i - r_i^{vdW}}{r_i^{vdW}} + \frac{d_e - r_e^{vdW}}{r_e^{vdW}}$$

As an example, we show in Figure 2 the Hirshfeld surface of catena-(bis(μ2-4-Nitrophenolato)-lead(II)) (CCDC identifier: FIRPOM) and its fingerprint plots. In Figure 2(a), the Hirshfeld surface color-coded by d_{norm} highlights the intermolecular interactions in crystal, which affect the assembly of molecular crystals into networks. The strong short range interactions are mapped out in red, whereas the porous regions are mapped out in blue, indicating long-range intermolecular interactions or even non-interactions. As mentioned previously, Hirshfeld surface (Figure 2(b)) can be color-coded with various properties, including d_i, d_e (Figure 2(c)), and their normalized distance (d_{norm}) by the Van der Waals radii of the involved atoms (Figure 2(d)). Despite being qualitatively visualized, the intermolecular interactions can also be captured quantitatively by the histogram of pairwise d_i and d_e sets. The Hirshfeld surface gives a unique signature of a molecule in a crystal, because it strongly depends on the surrounding, so the same molecule in different crystal packing looks different. Thus, the histogram of the geometric mappings to the molecule itself and its neighboring environment, is able to capture the high dimensionality of the descriptor spaces of materials databases, including chemical formula, molecular structures, crystal parameters, intermolecular interactions and *et al*.

The range of d_e and d_i across the Hirshfeld surface varies considerably depending on the atoms in the molecule (size dependence) and the particular type of intermolecular interaction experienced (interaction dependence). The ‘jet’ color map on the surface are customized to each group of molecules in order to present the contrast of contact distances. Whereas for d_{norm}, the diverging color map is used to illustrate the deviation from non-interaction state.

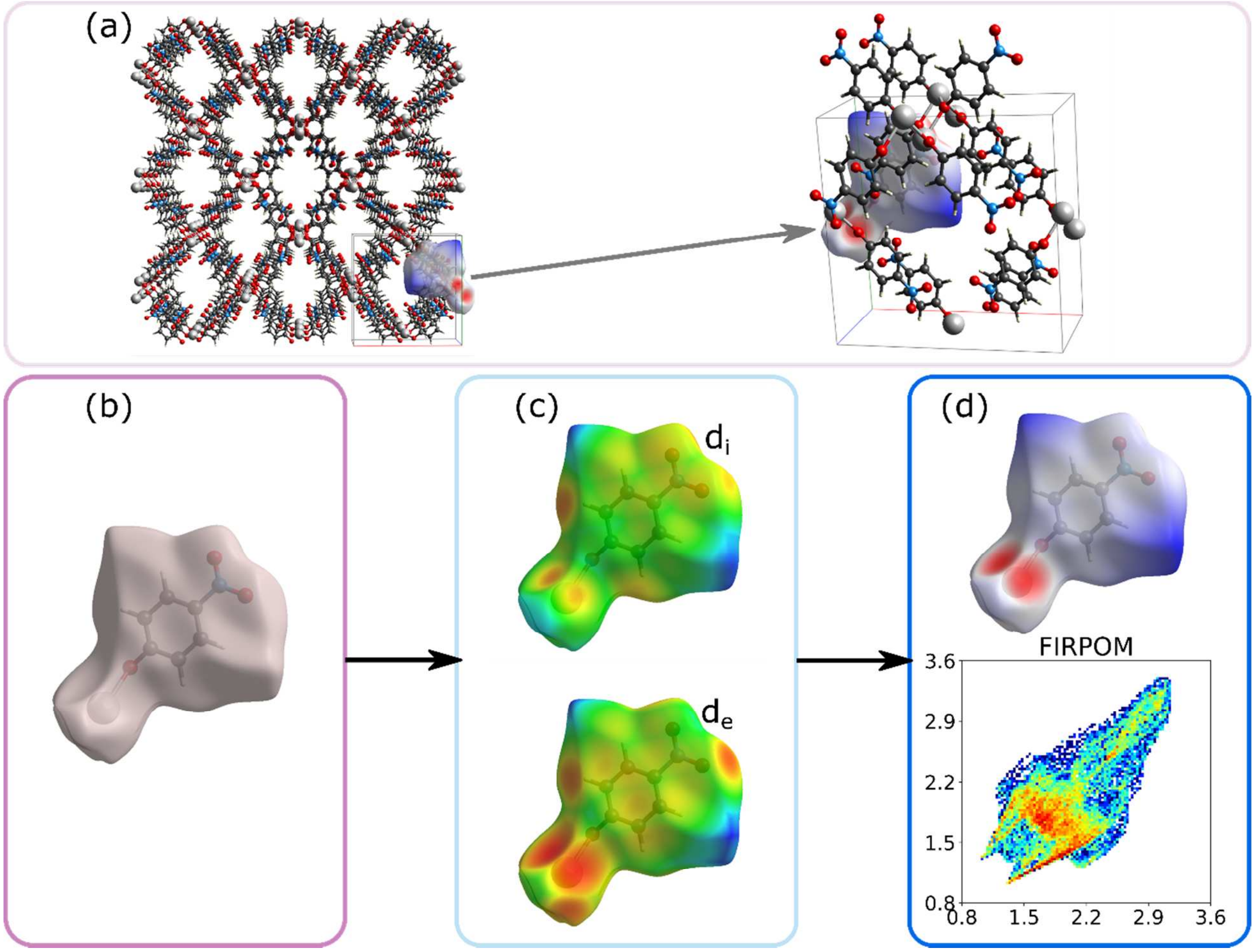

Figure 2 Finger-printing MOF structure (CCDC identifier: FIRPOM) and quantitative visualization of the molecule-in-crystal interactions. The Hirshfeld surface (c) of this molecule is encoded with a red-blue color scale reflected by the relative distance between pairwise d_i / d_e and the corrresponding Van der Waals radii. Red indicates close contact, where the interatomic distance is less than the Van der Waals radii, while blue indicates long range contact, where the interatomic distance is longer than the Van der Waals radii. (d) is the 2D histogram of d_i and d_e values for the example MOF structure.

Since Hirshfeld surfaces not only contain all of the geometric information of a compound but also the electronic structure and bonding information (involving solving spherical Schrödinger's equations of atoms), Hirshfeld surfaces and their fingerprint plots are able to discern differences that one may not be able to realize just through visual inspection of the MOF crystal structures. Despite the fact that the 2D fingerprint plot can visualize all close contacts and to decode each possible contribution involved within the structure, we need a sophisticated data/graph analysis tool to derive the quantitative structure-property relationships and the formation pathways for MOF structures. In this work, the outcome of

3D Hirshfeld surfaces are reduced to 2D histograms of d_i and d_e as the material fingerprints, which serve as the only input descriptors for the manifold learning algorithm mapping out the correlations.

2.2 Manifold Learning

In previous studies, we have successfully applied the use of the Isomap algorithm as a powerful tool to decipher the relationship between different chemistries[12, 13]. One of the features of the Isomap algorithm is that instead of assuming the data has an intrinsic linear structure, it allows the data, which is usually represented by a set of points, to be located on a non-linear manifold. Based on this allowance, it generates a low dimensional embedding of the data in a linear space (usually $\square^{2}$ or $\square^{3}$) such that the inner products between all of the pairs of data points preserves, with minimum error, the inner products between the data points on the original manifold.

This is done by first constructing a weighted graph (the neighborhood graph[13]), that has all of the data points $\{x_i\}$ as the vertices. The edges of this graph are determined by connecting each vertex to its k nearest neighbors in terms of the distance defined in the input space, with the weights of the edges as the input space distances between the two vertices that the edge is connecting to. By using the weighted graph, the geodesic distance can then be estimated as the least weighted path between an arbitrary pair of given points. With the estimated geodesic distances, the algorithm then applies the classical multidimensional scaling method to produce a low dimensional projection / embedding of the data that geometrically maps the correlations (in terms of the relative positions) between the points on the original manifold. Nevertheless, in practice the dimension of the manifold may be unknown, and for the purpose of visualization, the original data is commonly projected onto a 2D plane. The distance between two points in the low-dimensional projection may not be used as a measure for similarity, yet there is a one-to-one correspondence between the points on the original manifold and the points on the low-dimensional projection. The latter can be connected by edges that are determined in the first step to produce an unweighted graph network, which can be regarded as a similarity/dissimilarity map for the input points. Therefore, in general, two points in the input data are said to be similar if they are close in terms of Euclidean distance in the projected space and they are connected by a path in the graph network. For two points, the more edges in the shortest path that is connecting them, the less similar they are.

With the ultimate goal being to look for a new methodology for MOF design and selection with target properties based on the information / data that is already at our disposal, we applied the Isomap algorithm to the 2D fingerprint plots of the Hirshfeld surfaces as a first step towards this goal. Here, a MOF structure is represented by its 2D fingerprint plot which is treated as a point $\{x_i\}$ in a high dimensional Euclidean space. Since a 2D fingerprint plot is able to capture various properties, such as intermolecular interactions within the crystal structure, short / long-range contacts, and bonding information of the MOF compound, the points $\{x_i\}$ are assumed to lie on a manifold that contains all of this information, although the dimension of this manifold is not necessarily n. Based on the results from Tenenbaum *et al.*[13], it is reasonable to assume that the low-dimensional projection of the set of the 2D fingerprint plots will map some trend along the manifold in the input MOF structures, and the similarity / dissimilarity map will provide information about how different MOF chemistries and structures are related by using the connectivity in the graph.

The fingerprints of Hirshfeld surfaces of MOF structures can be thought of as points in a high-dimensional vector space (□ $^{500\times 500}$). The Isomap algorithm was then used to discover the optimal low dimensional graph which contains the information of the intermolecular interactions and bonding information of the MOFs in the input dataset, such that the geodesic distance relative geometric relations between the finger print of the MOFs in the higher dimensional manifold is preserved when mapped onto the reduced dimensional graph. The full details of the library of Hirshfeld surfaces and the treatment of the computational details of the manifold learning analysis are provided in the Supplementary Materials.

It is crucial to construct, in the first step, an appropriate graph so that the estimated geodesic distance between the vertices on the graph is an accurate approximation of the geodesic distances between points on the manifold. The effect of a “manifold short-circuit”[14] could occur if care is not taken for selecting the number of nearest neighbors when constructing the weighted graph in this step. To check for the potential short-circuits in the manifold, we compared the results by changing the number of nearest neighbors while looking at the residual variance of the reconstruction and the shape and connections of the resulting graph network.

3 Results and discussion

The 57 MOF structures investigated in this work are randomly chosen from the MOFs database[6] and have structures which are amenable for computation. Tthe Hirshfeld surface of those MOFs were generated and transformed into 2D histograms with 500 by 500 bins per graph.

Figure 3 presents the graph network (connectivity map) generated by the Isomap algorithm for these 57 MOF structures that we used as our template of study (see Supplementary Data for full library of calculated Hirshfeld surfaces and their fingerprints). The number of nearest neighbors and the number of reconstruction dimensions are two in this dataset for optimal mapping results (see section 3.1 and 3.2 in supplementary materials for the details on the optimization step). They represented over 30 different space groups, as we wanted to explore our machine learning approaches on as diverse set of crystal chemistries as possible. This manifold learning based graph network can be viewed as a form of structure map for reticular compounds: that is, an analogue to structure maps in crystal chemistry that seek patterns among diverse chemistries of compounds. The advantage of this network structure map is that it links chemistry and bonding to properties or other meso-scale attributes of the materials (in this case, surface area, which is a key metric in defining many physical properties of MOF structures). It should be noted that since the Hirshfeld surface is capturing a set of different types of information (chemical, physical and crystalline) about the MOF structures, so too does its 2D fingerprint plot. Thus, different edges in the graph network could represent different sets of information; e.g., the edge that is connecting MOF structures A and B represents the similarity in certain physical properties in A and B, yet the edge that is connecting MOF structures B and C represents some similarity in their chemical properties. As we will show in the following examples, different parts of the network structure map identify similarity in terms of the pore size between compounds of very different chemistries and crystal structure. The connectivity map draws attention to compounds that encourages one to ask what drives the apparent similarity. As discussed below, it can reveal structural features that are “hidden” and the data manifold derived from the Hirshfeld surfaces provides insight into the structural origins in the MOF structures that drives similarity in properties (in this case represented by surface area) different from structural similarities.

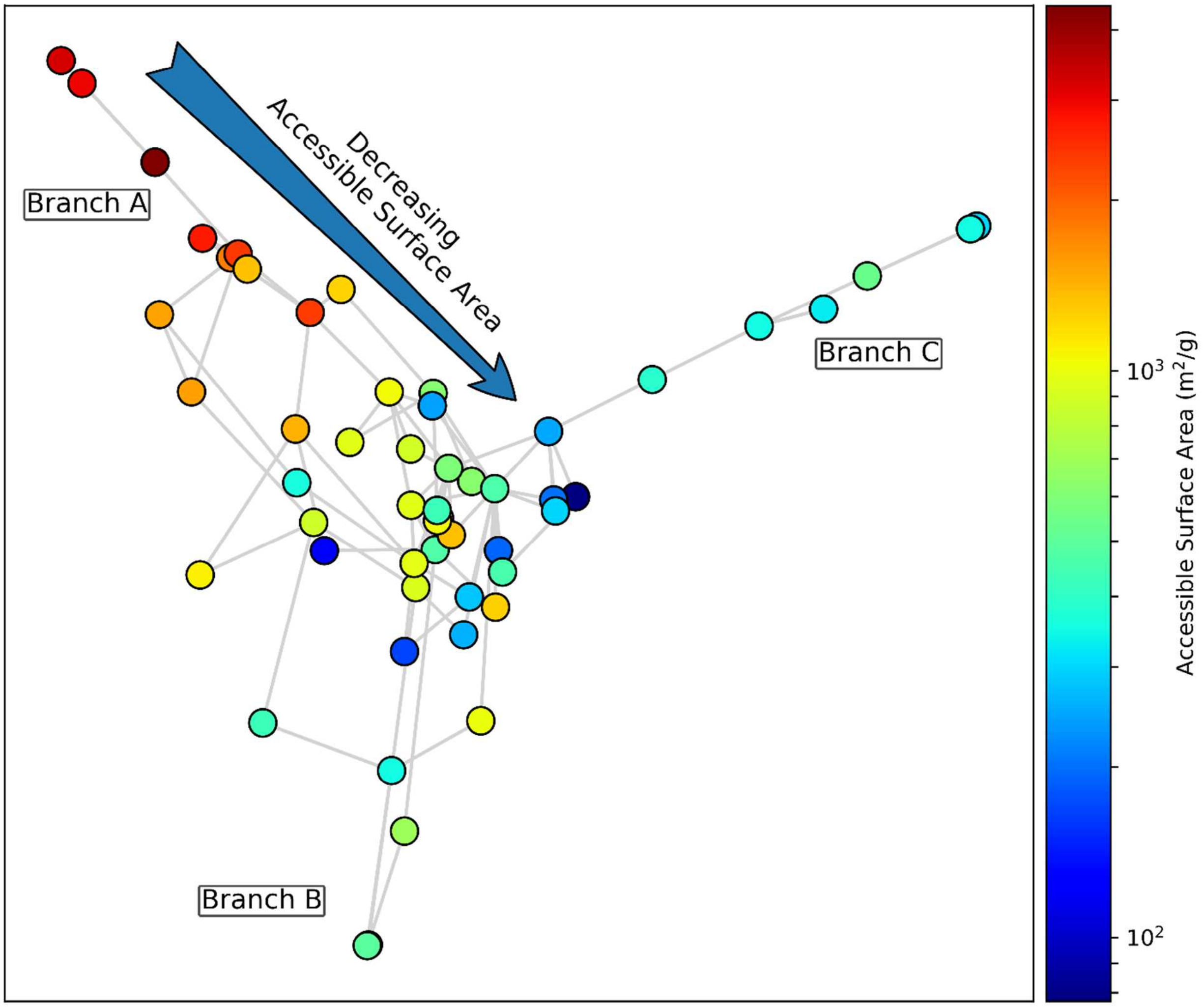

Figure 3 Graphical representation of mapping similarity/dissimilarity between MOF structures based on the 2D-fingerprint of their Hirshfeld surfaces. The vertices on the graphical map is color coded with accessible surface area. The lookup table of reference number, structure, and property information of MOFs can be found in Supplementary Data.

For this input set of MOF structures, the resulting perspective of the graph network can be qualitatively described in terms of an approximate 'Y' shape, with 3 branches and one centroid of cluster of compounds. To get a better insight of the extent of information captured by this graph network, the vertices are color-coded according to the corresponding accessible surface area, which is not an input of the Isomap algorithm. The colors of vertices show a globally decreasing trend of accessible surface area on the Branch A, as indicated by the blue arrow. For structures on Branch B and C, their accessible surface areas are approximately on the same order. Despite the fact that structures on branches B and C have similar orders of accessible surface area, the main difference between these two branches is the dimension of structures. That is, branch B has low dimensional MOFs, whereas branch C only has 3D MOFs. This global clustering could help the initial phase of MOFs design. For example, if we are searching for the low dimensional MOFs in the database, we can leave the materials on branch C out, thus narrowing down the search space. In addition to the global trend, Isomap algorithm can also serve as a visualization tool to capture the similarities between MOF structures, which are not easily discernable from simple inspection of either the crystal structure or Hirschfeld surface alone.

3.1 Exploring similarities between two- or one- dimensional MOF structures

One aspect of material similarity/dissimilarity is the range and strength of bonding / interactions, which has a significant impact on engineering layered materials. There are both strong covalent bonding and weak van der Waals interactions in those two- or one-dimensional layered materials[15]. Atoms are covalently bonded within the same plane, while the inter-planes interact with neighboring layers via van der Waals forces. With 2D fingerprint plots containing the information about the intermolecular interactions in the crystal, we are able to visualize the similarity between the 1D and 2D MOF structures in the input dataset.

Figure 4 highlights all of the 1D and 2D materials in the graph network, and their crystal structures are shown in Supplementary Material 1. In Figure 4, the 1D and 2D materials are mainly located in the center part of the network as well as on Branch A and C. On Branch A, Isomap finds some correlation between vertices 29 and 55 (CCDC identifiers / space group are LILWAE / I m m a and ZBPPHN01 / P 2_1 2_1 2_1, respectively) by showing a close connecting edge among all the input points (see Supplementary Material for zoomed-in figures). They share the same metal ion in the secondary building unit (SBU) in 1D layered structures as illustrated in Figure 5. This correlation is not easily to discover by just looking at the large database with sparse descriptors. Not to mention that they have different chemical formulas and crystal parameters.

Furthermore, the clustering on the end of Branch C are all 2D materials (vertices 8, 25 and 56 with CCDC identifiers being DEGJEG, KELJIV and ZIVDIT, respectively). Furthermore, the distance between vertices 8 and 56 is closer than that of vertex 25, indicating structure 8 and 56 are more closely related, with the same chemical formula $(C_6H_4Cu_1N_5)_n$ and space group of P 2_1 / c in Figure 6. Thus our method has a certain level of flexibility and is capable of detecting subtle differences between different compounds.

In the paper introducing Isomap, it is obvious to associate the features with the two detected dimensions of face and fingers[13]. However, for the ‘fingerprint’ of MOFs, the fingerprinting process itself has lost some extent of the information, although it has retained and summarized most of descriptor space. Thus, the correlation between the MOFs structures to the two detected dimensions is difficult.

In this study, the data lies along a broad curve in the projected space, which appears to trace the accessible surface area of the MOF structures. Although this methodology may not be feasible for the task of classifying MOF structures, it does help in getting an insight and big picture of the database of MOFs. It could also give information about how to move forward with the current state-of-the-art large and sparse database, such as how to preprocess the data before building a classification pipeline in the initial phase of MOFs design project.

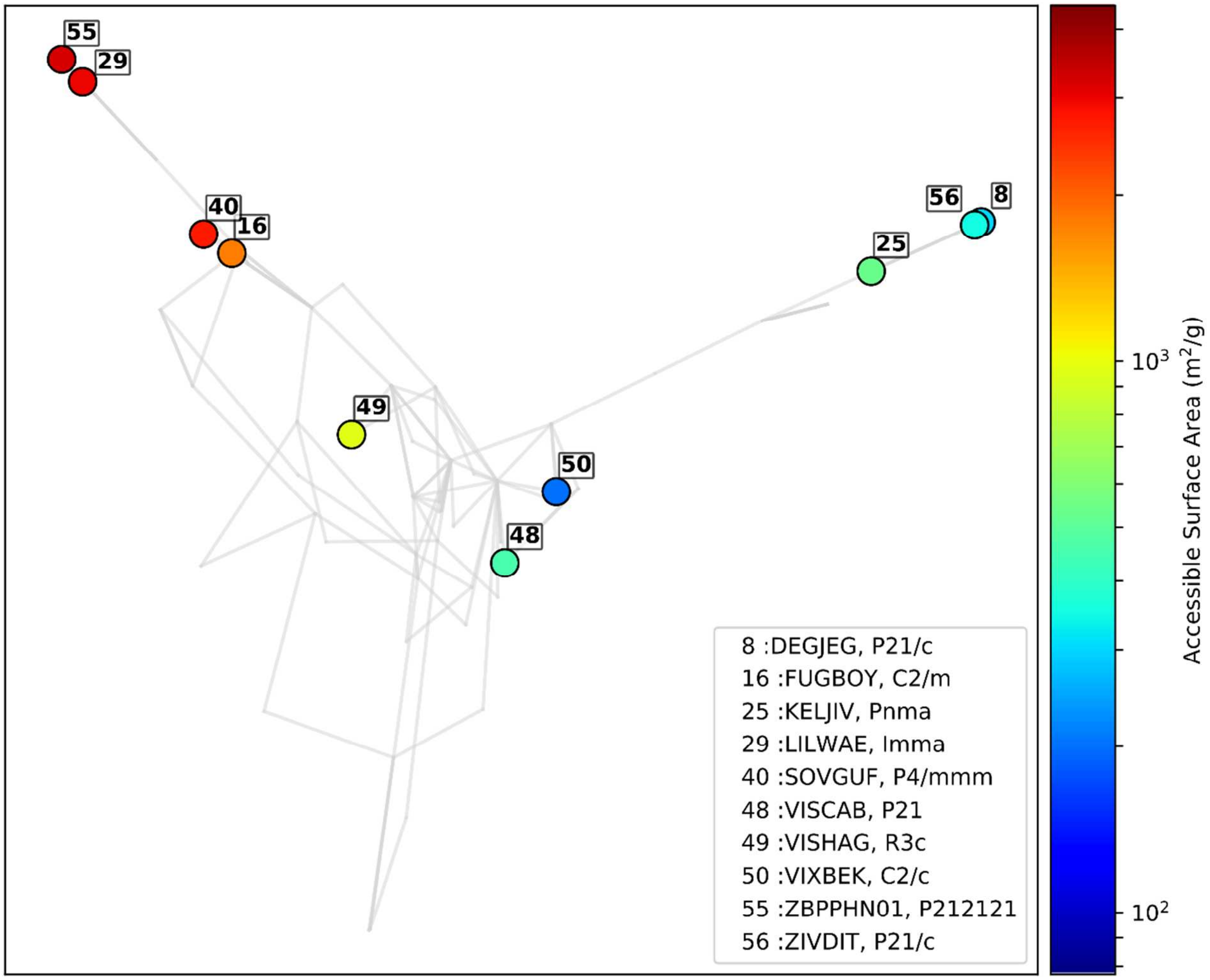

Figure 4 Two- and one-dimensional weakly bonded MOF structures with their correlations visualized by a graph network

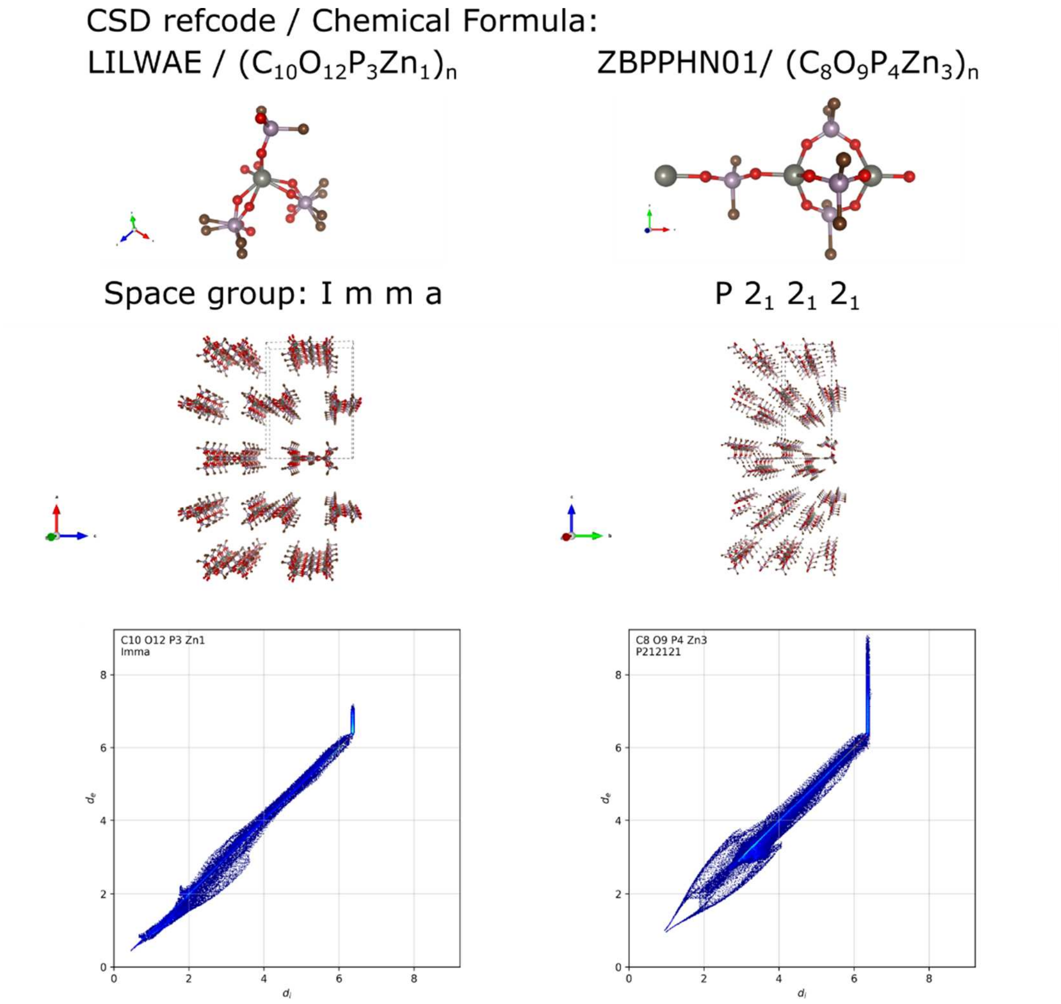
CSD refcode / Chemical Formula:
LILWAE / $(C_{10}O_{12}P_3Zn_1)_n$
ZBPPHN01/ $(C_8O_9P_4Zn_3)_n$
Space group: I m m a
$P\ 2_1\ 2_1\ 2_1$
C10 O12 P3 Zn1
Imma
C8 O9 P4 Zn3
P212121
0
2
4
6
8
d_e
d_i

Figure 5 Case studies of molecular, crystal and fingerprint plots of 1D MOF structures

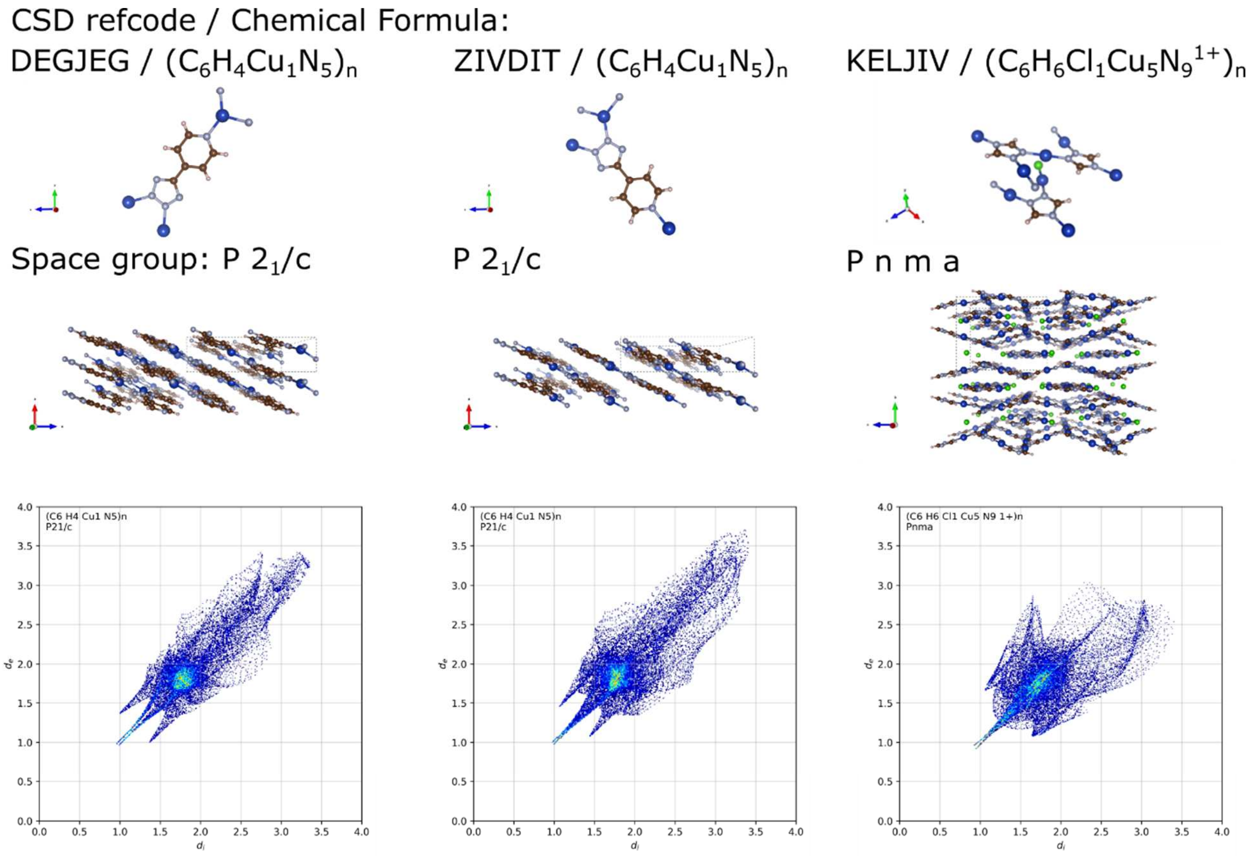

Figure 6 Case studies of molecular, crystal and fingerprint plots of 2D MOF structures

3.2 Identifying Compounds with Similar Building Unit Coordination Chemistry

The two MOF structures 13 and 31 on Branch B have one single edge connecting them, and by inspection we found that they have the same coordination unit, see Figure 6. Therefore, this part of the network structure map is capturing MOF structures having the same coordination unit. The materials on the first two vertices are 2D and 2D/1D materials with relatively comparable accessible surface area and with pore-geometries featuring square cross-sections of the channels. From FIRTEH to MIJSII, their crystal structures and pore geometries are not similar. The crystal space groups for FIRTEH and MIJSII are $P\bar{3}m1$ and $P4/nmm$, respectively. However, the Isomap algorithm finds their similarities manifested in the secondary building units (SBU). Namely, both of their SBU consist of a binuclear unit of distorted octahedral CuO_2N_4 sites with the $Cu_2(azole)_2$ planar core, as shown in Figure 8. The coordination geometry at each Cu(II) site is defined by the oxygen donors of a chelating acetylacetonate ligand, two tetrazolate nitrogen donors and two pyridyl nitrogen donors (Figure 9). Since the HS captures both molecular geometry and packing characteristics of those molecular building units, we can find structural features that are unique in pore design.

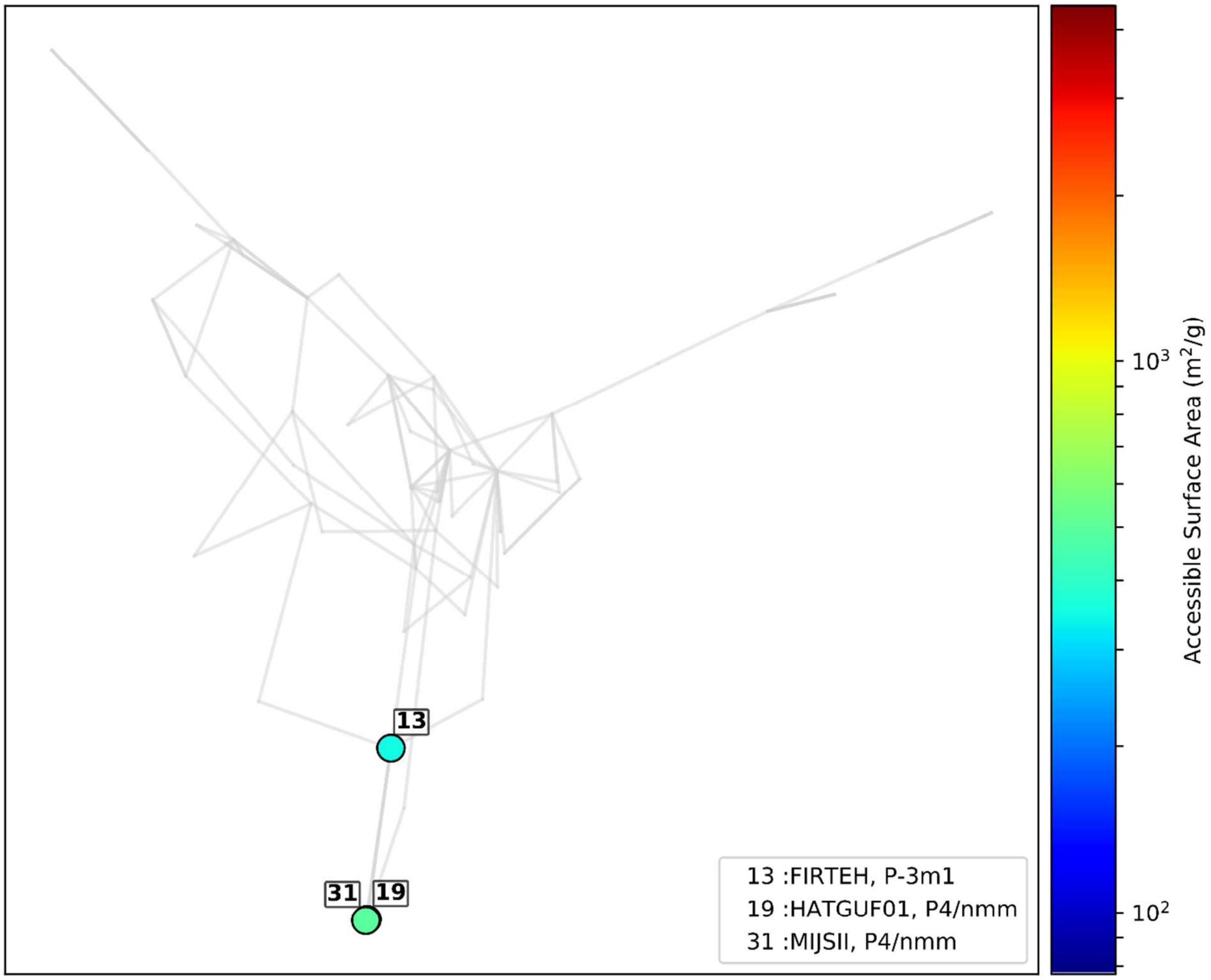

Figure 7 Mapping similarities of site chemistry between structure 31 (CCDC identifier: MIJSII) and structure 13 (CCDC identifier: FIRTEH)

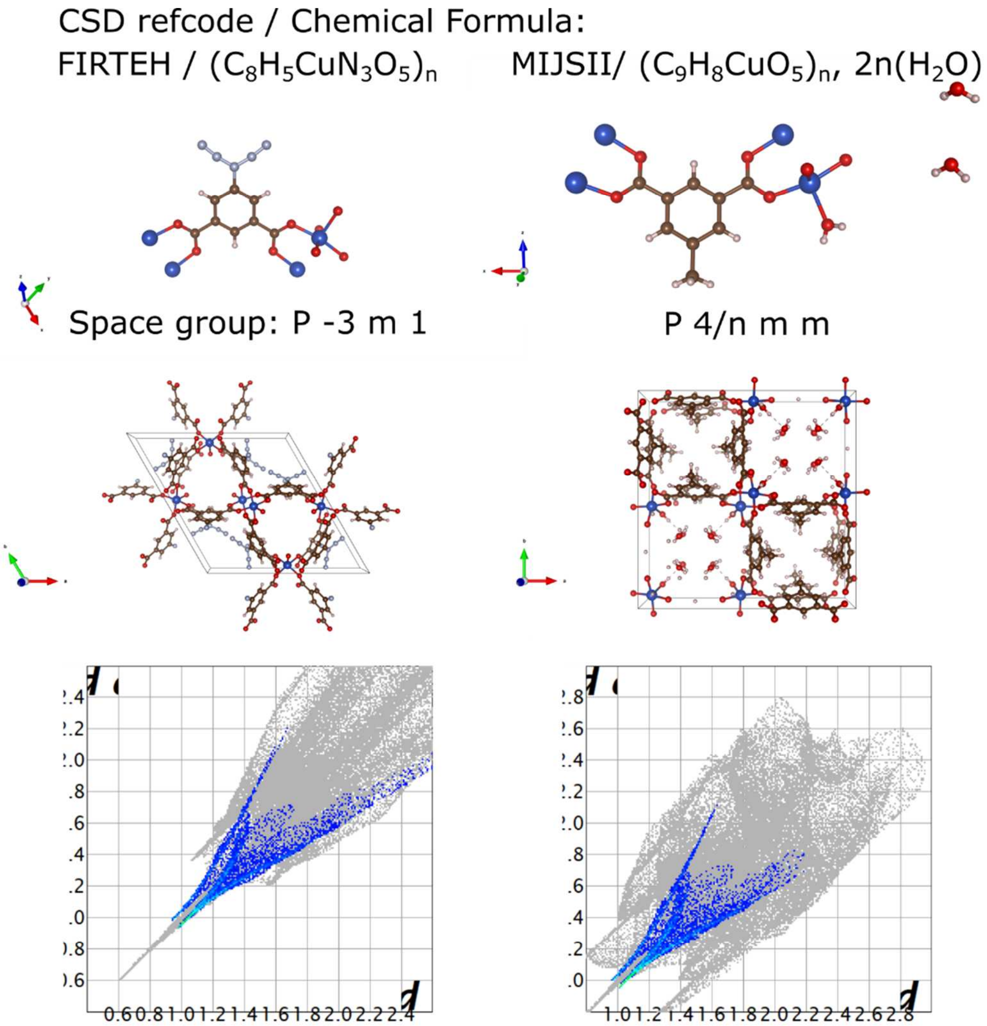

Figure 8 Case studies of molecular, crystal and fingerprint plots of structures with similar site chemistry.

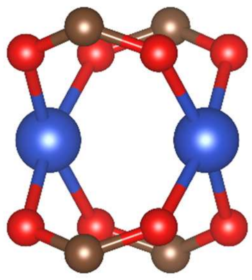

Figure 9 Metal-containing SBU of FIRTEH and MIJSII structures Cu2(CO2)4. Atom colors in molecular drawings: C, ochre; O, red; Cu, blue squares.

3.3 Identifying Leading Compounds for MOF design

A fundamental question in discovering and designing new molecules is to know where to start. Identifying a template structure to lead the search is of course a well-established field in bioinformatics and drug discovery. The connectivity map can serve as a foundation for identifying such leading compounds in designing MOF structures.

Here we propose the idea of a 'lead MOF Compound' ('lead' not 'Pb' compound!), borrowing the concept from the drug discovery field, for the set of input MOF structures. Specifically, for a graph network generated from a given set of MOF structures, we can identify the vertices with a relatively large number of edges connected to it. From a graph theoretic point of view, these vertices will have relatively high centrality in the graph and thus can be considered as containing more information than others about the graph network. From the similarity point of view, the existence of an edge between two vertices generally means that there is correlation between the two MOF structures; thus the compounds with a relatively large number of edges connected would potentially have properties that are in common with more of the MOF structures in the dataset than others. Starting from a lead MOF compound vertex, one can trace the edges to reach other MOF structures, and this provides a design path with potential information on design rules.

In our case, we choose the vertex in the graph network with the most edges connected to it: namely, $(C_{11}H_9N_7O_2Zn_1)_n$ with space group P n a 2_1 , which is shown as vertex number 42 (CCDC identifier: TEPGUS) in Figure 10. Also shown in Figure 8 are the vertices which are connected to vertex 42 by one edge and the corresponding crystal structures. In total, there are 17 MOF structure that are considered as similar to TEPGUS, and from the crystal structure we can see there are 11 different space groups and that the pore geometry varies significantly. This diversity indicates the nature of a lead compound. The crystal structure of TEPGUS is shown in Figure 11, which features the interconnected channels. By viewing the cross-sections from three different major crystallographic axes, significant channels are identified in all these directions with qualitatively similar cross sections geometries. This may suggest that $(C_{11}H_9N_7O_2Zn_1)_n$ has a structure that inherently accommodates a diversity of trajectories of channels and thus can serve as a signature compound from which one can build and explore other chemistries for targeted surface areas. One of the physical properties of 'lead compound' discovered here features high CO_2 uptake at 0.1 bar (273 K: 73 mg g^{-1}; 298 K: 49 mg g^{-1}) and high enthalpy of CO_2 adsorption (40 kJ mol^{-1})[16]. This is due to the rich hydrogen donors on the framework, which are in favor of the adsorption of the guest molecules into the pores. Although the 'lead compound' has characterized to be useful for gas adsorptions, it may nevertheless have suboptimal structure that requires modification to fit better to the specific guest molecules. We hope the methodology presented in this work, as well as the concept of 'lead compound', could help accelerate the discovery of novel MOFs and tuning the gas adsorption performance in the community.

In this study, we have used pore size and geometry as a metric for designing MOF structures and the value of using our manifold learning approaches to identify candidate systems as lead compounds to initiate a discovery process. This approach can have impact in many areas of study including the study of guest-host interactions in MOF structures. Guest-host interactions in MOFs is a multiscale problem and ascertaining the complexity of what combination of crystallographic and chemical features in MOFs is a challenge[17, 18]. The strategy proposed here provides a framework for identifying candidate compounds around which MOFs can be engineered to be tailored for specific guest-host interaction. This is one of the areas of further study we will be embarking on for future reports.

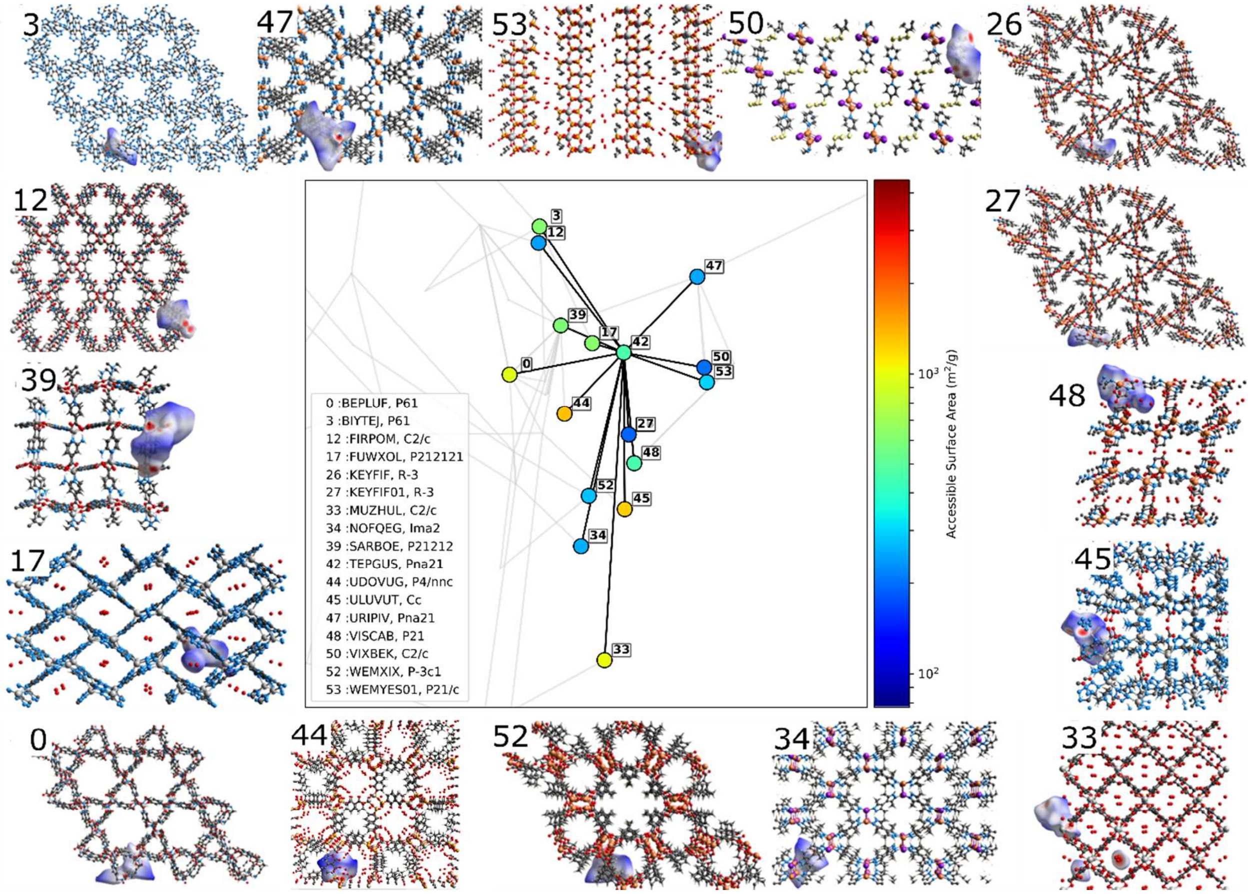

Figure 10 'lead compound' (structure 42, chemical formula: $(C_{11}\ H_9\ N_7\ O_2\ Zn_1)_n$) with its connected/similar compounds

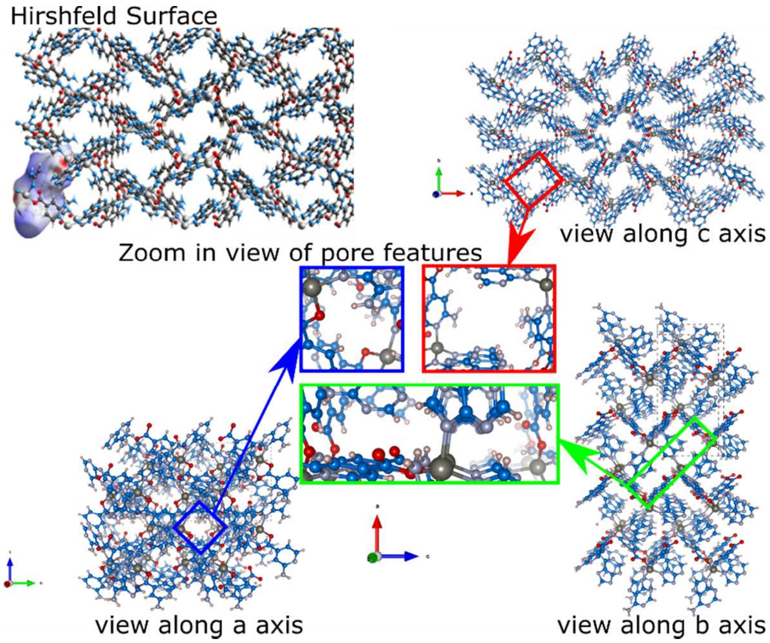

Figure 11 Hirshfeld surfaces, and perspective view along the three major crystallographic axes of the unit cell of the 'lead compound'

3.4 Scalability and Reproducibility

In this section, we expand our exploration space by an order of magnitude from 57 to 508 compounds. The resulting high-throughput calculations and the transformed Hirshfeld surfaces of MOFs followed the same mathematical procedures and the systematic statistical analysis previously described. Due to the expansion in information space, the 2D recoveries of the manifold are not sufficient to represent the embedded correlations (see section 6.3 in supplementary materials for detailed discussion), and therefore we extend the graphical networks to the first three detected dimensions in Figure 12. This larger network shows that the global trend tracing the accessible surface area is indeed valid with the extended dataset (for better visualization, please refer to the animated 3D network in supplementary document). Our findings regarding the detection of the significance of the specific structural building units and the identification of unique features associated with 1D and 2D structures are preserved from section 3.1 and 3.2 (see section 6.2 in supplementary materials for detailed discussion). We found a new set of compounds that serve as 'lead' compounds, but that in itself is not surprising given that we have far more compounds to choose from.

In order to test the performance and stability of this approach, we performed manifold learning on increasing sample sizes with random sampling drawn from the 508 populations. There are 10 experiments for each case, ensuring the reproducibility and robustness of the parameterized study. The statistical sampling results show that in order to minimize the reconstruction error with an increasing number of MOFs, we need a larger number of nearest neighbors (6 for this dataset) and higher dimensionality. With the tuned parameters, we have reproduced the same observations as described above but with a data set that is 10 times larger.

Thus, the scalability of our method is demonstrated without sacrificing the performance.

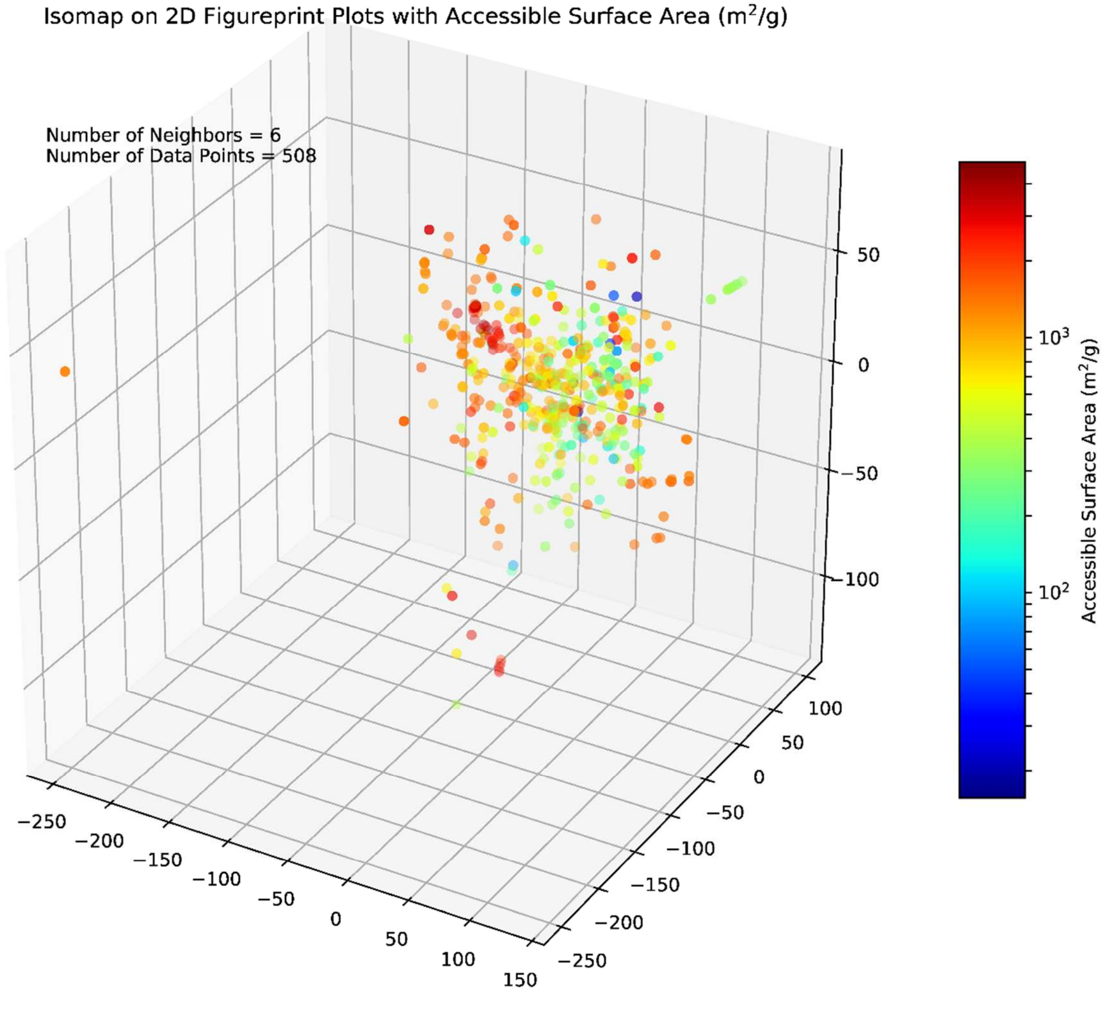

Figure 12 3 dimensional visualizations of the similarity/dissimilarity between 508 MOF structures based on the 2D-fingerprint of their Hirshfeld surfaces. The vertices on the graphical map is color coded with accessible surface area. The lookup table of reference number, structure, and property information of MOFs can be found in supplementary spreadsheets.

4 Conclusion

In this paper we have provided a new template for interrogating MOF structure databases. Harnessing the information in crystallographic databases to compute Hirschfeld surfaces, we have created a new library of information that fuses both structural and chemical bonding information. Manifold learning provides a means to navigate and interrogate this new data space to reveal new structure-property correlations that would not have been discernible by just looking at either electronic structure or crystallographic structure alone. The example provided in this paper used a limited set of compounds to demonstrate the technique but is readily scalable to much larger databases. We see this as the foundation for a database tool that can leverage the rapid growth of data in MOF structures derived from experiments and simulations.

Acknowledgements

We gratefully acknowledge support from National Science Foundation (NSF) DIBBs program, award number 1640867. The authors would also like to acknowledge support from the Toyota Research Institute Accelerated Materials Design and Discovery program. KR also acknowledges the Erich Bloch Endowed Chair at the University at Buffalo- State University of New York. We gratefully acknowledge support from the Collaboratory for a Regenerative Economy (CoRE).

Reference

1. J. Jiang, Y. Zhao and O. M. Yaghi, *J Am Chem Soc*, 2016, **138**, 3255-3265.
2. P. Z. Moghadam, A. Li, S. B. Wiggin, A. Tao, A. G. P. Maloney, P. A. Wood, S. C. Ward and D. Fairen-Jimenez, *Chemistry of Materials*, 2017, **29**, 2618-2625.
3. K. E. Cordova and O. M. Yaghi, *Materials Chemistry Frontiers*, 2017, **1**, 1304-1309.
4. C. S. Diercks, Y. Liu, K. E. Cordova and O. M. Yaghi, *Nat Mater*, 2018, DOI: 10.1038/s41563-018-0033-5.
5. M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib and R. Srivastava, *ACS Comb Sci*, 2017, **19**, 640-645.

6. Y. G. Chung, J. Camp, M. Haranczyk, B. J. Sikora, W. Bury, V. Krungleviciute, T. Yildirim, O. K. Farha, D. S. Sholl and R. Q. Snurr, *Chemistry of Materials*, 2014, **26**, 6185-6192.
7. M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji and T. K. Woo, *J Phys Chem Lett*, 2014, **5**, 3056-3060.
8. H. Ghasempour, A. Azhdari Tehrani, A. Morsali, J. Wang and P. C. Junk, *CrystEngComm*, 2016, **18**, 2463-2468.
9. K. Rissanen, *Chem Soc Rev*, 2017, **46**, 2638-2648.
10. S. K. Seth, A. Bauzá and A. Frontera, *CrystEngComm*, 2018, **20**, 746-754.
11. G. Mahmoudi, H. Chowdhury, B. K. Ghosh, S. E. Lofland and W. Maniukiewicz, *Journal of Molecular Structure*, 2018, **1160**, 368-374.
12. S. Srinivasan, S. R. Broderick, R. Zhang, A. Mishra, S. B. Sinnott, S. K. Saxena, J. M. LeBeau and K. Rajan, *Sci Rep*, 2015, **5**, 17960.
13. J. B. Tenenbaum, V. De Silva and J. C. Langford, *science*, 2000, **290**, 2319-2323.
14. M. Balasubramanian and E. L. Schwartz, *Science*, 2002, **295**, 7-7.
15. G. Cheon, K. N. Duerloo, A. D. Sendek, C. Porter, Y. Chen and E. J. Reed, *Nano Lett*, 2017, **17**, 1915-1923.
16. E. Yang, H.-Y. Li, F. Wang, H. Yang and J. Zhang, *CrystEngComm*, 2013, **15**, 658-661.
17. S. Amirjalayer and R. Schmid, *The Journal of Physical Chemistry C*, 2016, **120**, 27319-27327.
18. Z. Li, Z. Zhang, Y. Ye, K. Cai, F. Du, H. Zeng, J. Tao, Q. Lin, Y. Zheng and S. Xiang, *Journal of Materials Chemistry A*, 2017, **5**, 7816-7824.