MRS Communications (2021) 11:811–817

https://doi.org/10.1557/s43579-021-00097-x

MRS Communications

Research Letter

Exploring the shape of data for discovering patterns in crystal chemistry

Scott Broderick, Ruhil Dongol, and **Krishna Rajan,** Department of Materials Design and Innovation, University at Buffalo, Buffalo, NY, USA
Address all correspondence to Krishna Rajan at **krajan3@buffalo.edu**

(Received 30 June 2021; accepted 13 September 2021; published online: 23 September 2021)

Abstract

This paper describes an unsupervised exploratory data mining strategy for identifying significant chemistry-structure-property relationships in complex crystal chemistries. Using the formalism of Topological Data Analysis, (TDA) we show how hierarchical patterns in families of crystal structures of the apatite family of the type $A1_5A2_5B1B2O_{24}X$ can be automatically detected via TDA. This 'self-driving' data exploration approach is shown not only to uncover links between structural building units and different stoichiometries but also to uncover new and yet unexplored associations between properties and coordination polyhedral geometry.

Introduction

"Is there some pattern to these complex crystal structures that eludes the casual observer, but which can explain essential features of their structures?" [1-2]. The pioneering crystallographer Alan Mackay has expounded on the idea of a framework for 'Generalized Crystallography' for over half a century [3-6], to exploit the role of crystallography from one of classifying structures to harness that knowledge to one that uncovers promising new structure-chemistry-property relationships in chemically and structurally complex materials [6-7].

Mackay has proposed that 'the crystal is a structure, the description of which is much smaller than the structure itself' and that this description of structure serves as a 'carrier of information' about the structure on larger length scales. The many crystal chemistry design rules such as the classical Pauling rules and others manifest themselves on their influence on crystal structure in terms of the local coordination environment and the resulting packing arrangements of the symmetrical coordination polyhedra. As noted by Mackay, these rules are very successful in describing the immediate local environment but are not so easily extended to less symmetrical polyhedra and to complex crystal chemistries that are hierarchic in nature [8]. The importance of the concerns raised by Mackay lies in the fact that there may be some pattern in complex crystal structures that have eluded discovery but can explain essential characteristics or properties of such structures. Knowing this opens the door to a materials science exploration of looking beyond existing framework of knowledge and hence to the foundation for exploratory data mining in an unsupervised fashion. In this paper, we explore this issue in more detail by the application of the new advancements in harnessing the mathematical formalism of topological data analysis. We show in this prospective article how given the advances in materials informatics in the last two decades, Mackay's vision of crystallography can be realized and serves as a blueprint for materials discovery and design.

The fundamental challenge has been to integrate the 3-dimensional metrics defining geometric complexity with the n-dimensional sets of metrics that relate to chemical bonding. The latter includes but is not limited to attributes associated with individual elements, pair wise bond interactions, quantum level information at both the local atomic environment level and packing characteristics of fundamental structural units to form the unit cell, to mention just a few [9-11]. From this point cloud of information, the challenge is to seek and detect patterns and connections with the least amount of bias to guide that process. From examining all possible patterns, we can then explore via inference

potentially significant relationships that would otherwise go undetected. The mathematical discipline of topology provides a quantitative framework to address this challenge by acting as a "microscope" where one can observe and record high dimensional data sets from different perspectives, "magnifications" and "focus", all of which can be manipulated independently [12]. As summarized by Carlsson [13], Toplogical Data Analysis (TDA) can be viewed as an intermediate methodology between modeling and cluster analysis. The former is governed by algebraic equations that are continuous but not very flexible, while cluster analysis is discrete and therefore misses continuous phenomena (shapes) captured in the output of TDA.

In a recent study we introduced the use of topological data analysis (TDA) as an unsupervised machine learning tool to uncover classification criteria in complex inorganic crystal chemistries [14]. Using the apatite family of structures as a template that contains all the complexities of low symmetry coordination polyhedra and hierarchical structure in crystal chemistries, we tracked through the use of persistent homology the topological connectivity of input crystal chemistry descriptors on defining similarity between different stoichiometries of apatites. It was shown that TDA automatically identifies a hierarchical classification scheme within apatites based on the commonality of the number of discrete coordination polyhedra that constitute the structural building units common among the compounds. We provide an alternate data driven approach of objectively classifying datasets using concepts from algebraic topology, namely, persistent homology. In this paper, we expand on how TDA identified key bond pairs that were common between compounds and when we outline those bond pairs, we find that the TDA was identifying the connections that outlined the basic structural building unit/polyhedra. We also demonstrate how this can serve as a platform to identify unexplored structure-chemistry-property associations.

Methods

Within the TDA framework [16-19], persistent homology (PH) is the most commonly used method and captures the topological features in the point cloud. Here, the point cloud is the high-dimensional chemical space of the apatite family, where each data point (component) represents an apatite compound [14,20]. In practice, PH does not directly study the point cloud but maps a set X points in the high-dimensional space associated with a distance function, in our case the Euclidean distance function. The mapping operates by placing a ε (filtration) -radius disc centered at each component in X to form an overlapped space $\tau_\varepsilon(X)$, which defines the set of all points within the disc. A connection between two components is established when the topological space occupied by their discs overlap. The number of connected components in $\tau_\varepsilon(X)$ increases with larger ε, while for sufficiently large ε, all components are connected. Naturally, the topological features evolve with increasing ε; hence, the fundamental idea behind PH is the study of the evolution of the topological features in $\tau_\varepsilon(X)$ as ε increases.

The evolution of connectivity of the topological features in $\tau_\varepsilon(X)$ along the filtration value is modeled using a sequence of nested geometrical constructs as defined by the Vietoris-Rips simplicial complex. These are tracked and visualized through a series of connectivity diagrams and compressed in a static figure called a persistence barcode. In this work, the barcode only represents the zero-order homology group (H_0, path-connected components) containing information about the evolution of data connectivity. Higher-order homology groups are discarded as they do not contain useful information for this analysis. The persistence barcode provides a quantitative description of the chemical space by recording the birth (initial connection between new data points) and death (combining the two data points) of a topological feature. For connectivity analysis, the birth of barcodes for all components occur at ε=0 and the death (termination) of a barcode signifies formation of a connection between two components. Figure 1 summarizes the algorithmic workflow from the point cloud representation to the coordination driven interpretation extracted from the barcodes. In essence, the PH transforms the chemical space into a set of barcodes that captures nuanced chemical information through topological constructs.

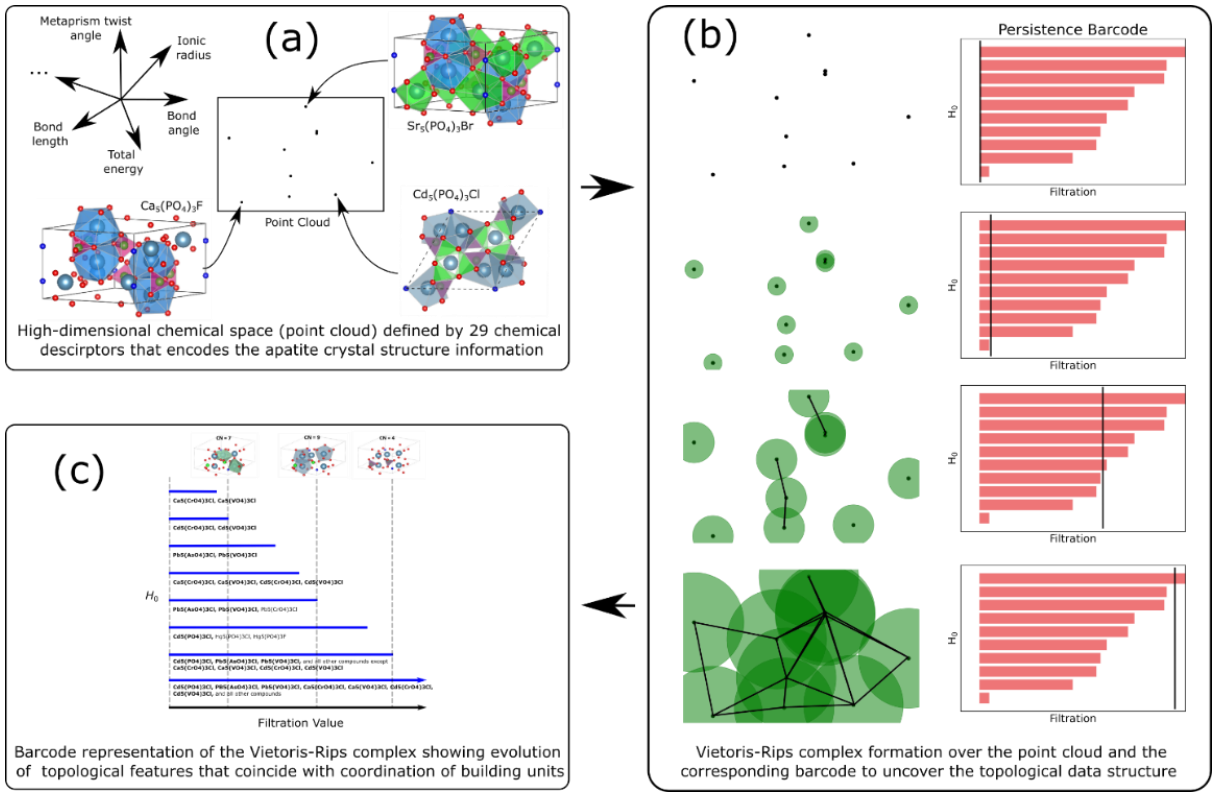

Fig. 1 The algorithmic workflow from point cloud representation to coordination drives interpretation of the persistence barcode. (a) The point cloud is a high-dimensional chemical space defined by the descriptor set [14,20] that encodes the crystal structure information. (b) Performing Vietoris-Rips complex calculation over the point cloud results in connection between data points driven by the overlapped space of $\tau_\varepsilon(X)$ for a given ε (filtration). (c) The evolution of the barcodes coincides with coordination driven building units (recreated from our previous work [14]). PH transforms the chemical space into a set of barcodes that capture nuanced chemical information through topological constructs.

Results and Discussion

In our recent work on classifying apatites [14], we linked chemistry of the compounds and the persistence of their features with their coordination polyhedra. The similarity between chemistries is greatest when they are grouped at lower filtration values. Specifically, the barcode tells us which aspects of stereochemistry dominate in the building of apatite structures, from which we discovered a hierarchy of dominance of coordination numbers for families of apatites. This tool allows us to explore alternate chemistries and crystallographic parameters, which can be linked to material properties. TDA provides a representation of data in the form of a barcode. To aid in interpretation, we converted the barcode to a dendrogram [21] and here we now expand that to property, and specifically in this paper we focus on bandgap. The bandgap values were extracted from Materials Project [22] and AFLOW [23]. As highlighted in Figure 2, as the filtration value is increased and the bars are growing, more compounds are classified together. This classification is done based on the connectivity of the data.

From the collection of barcodes, we extract the filtration value which corresponds with each joining of compounds. For example, $Ca_5(PO_4)_3F$ and $Sr_5(PO_4)_3F$ are originally on their own bars, but then join bars at filtration value ε of 0.2. We can then convert that to a dendrogram by having the vertical line connecting the compound lines at $\varepsilon = 0.2$. As ε continues to increase, more compounds join the bar. At $\varepsilon = 0.6$, $Ba_5(PO_4)_3F$ joins $Ca_5(PO_4)_3F$ and $Sr_5(PO_4)_3F$ in the

barcode, and then accordingly we have a vertical line at ε = 0.6 in the dendrogram connecting these components of the dendrogram. This approach continues until we have the entire barcode represented as a dendrogram.

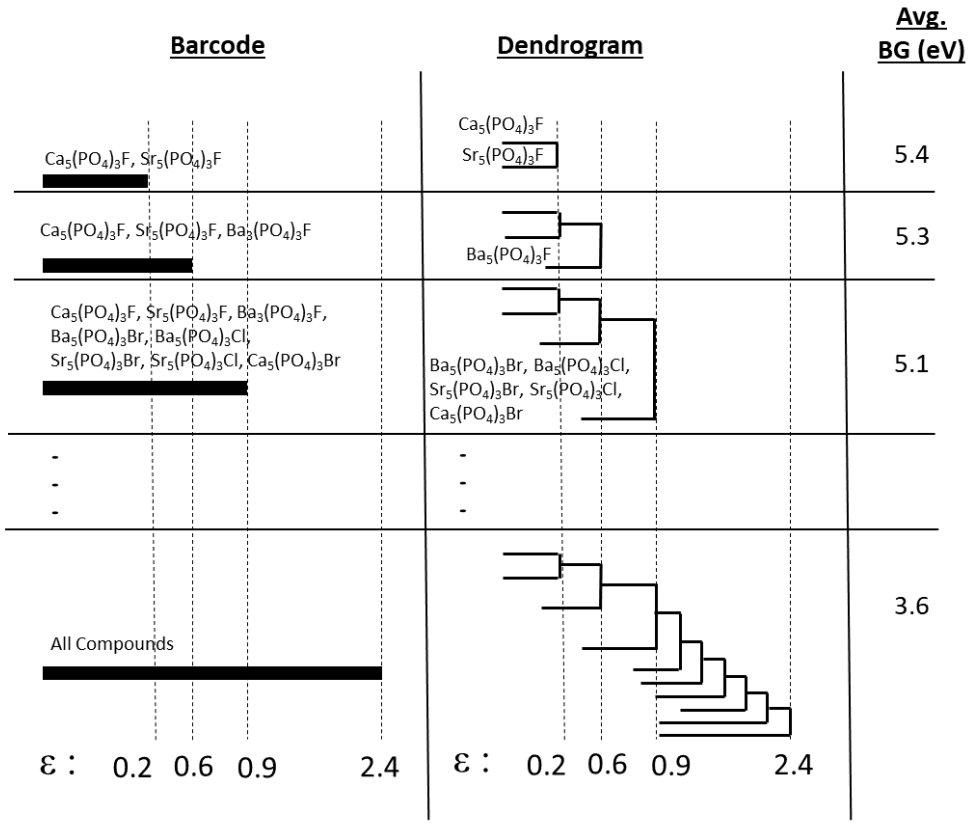

Fig. 2 Relationship between barcode and dendrogram representation, as well as property, which is this case is the bandgap of the compounds at each level. As ε increases, more compounds start to merge onto the same bars. These are compounds with higher similarity, but this merging can also be represented as a dendrogram with the corresponding filtration values. Interestingly, as the similarity with the starting compounds decreases, the bandgap also decreases. This allows us to understand the relationship between crystal geometry, atomic coordination and properties through the application of TDA.

The barcode provides a framework to create a classification of compounds, when the classifiers do not have any obvious relationships. By looking at the apatite chemistries, we are able to find groupings of the materials in terms of the polyhedral make-up. Our technique captures the dominancy of certain coordination number and site occupancies. Through this, we are able to guide where we change our design features by defining specific site chemistries with given coordination numbers to adjust material properties. The static similarity metric can be extracted from the barcode by taking snapshots at various filtration values. In a traditional 2D structure map, the closer two compounds are than the more similar they are. Generalizing this idea to the snapshots of the barcode, the sooner (ie lower filtration value) two compounds merge (to appear on the same bar), the higher similarity they have. In the extreme, we can think of it as two compounds with identical descriptors appear at the same point in a 2D structure map or share the same bar in a barcode no matter the filtration parameter. This then provides a framework for chemical substitution with minimal impact on properties, i.e., shorter bars indicate that the compounds on it have more similarity in their

properties. Following this logic, we could apply this information to suggest when it is possible to replace toxic or undesirable elements.

Property can also be overlaid with this information. For example, the average bandgap of $Ca_5(PO_4)_3F$ and $Sr_5(PO_4)_3F$ is 5.4 eV. When we add $Ba_5(PO_4)_3F$ to the bar, the average bandgap drops to 5.3 eV. As more compounds are added to the bar with the increasing filtration value, we find that the average bandgap continues to drop, to the final average value 3.6 eV. This result introduces an interesting added aspect to the use of TDA, showing for the first time the correlation between a property which was not input in any manner into the analysis and the resulting material barcode. Thus, the connectivity of data represents the change in apatite bandgaps. The variation in bandgap with the TDA result is highlighted in Figure 3.

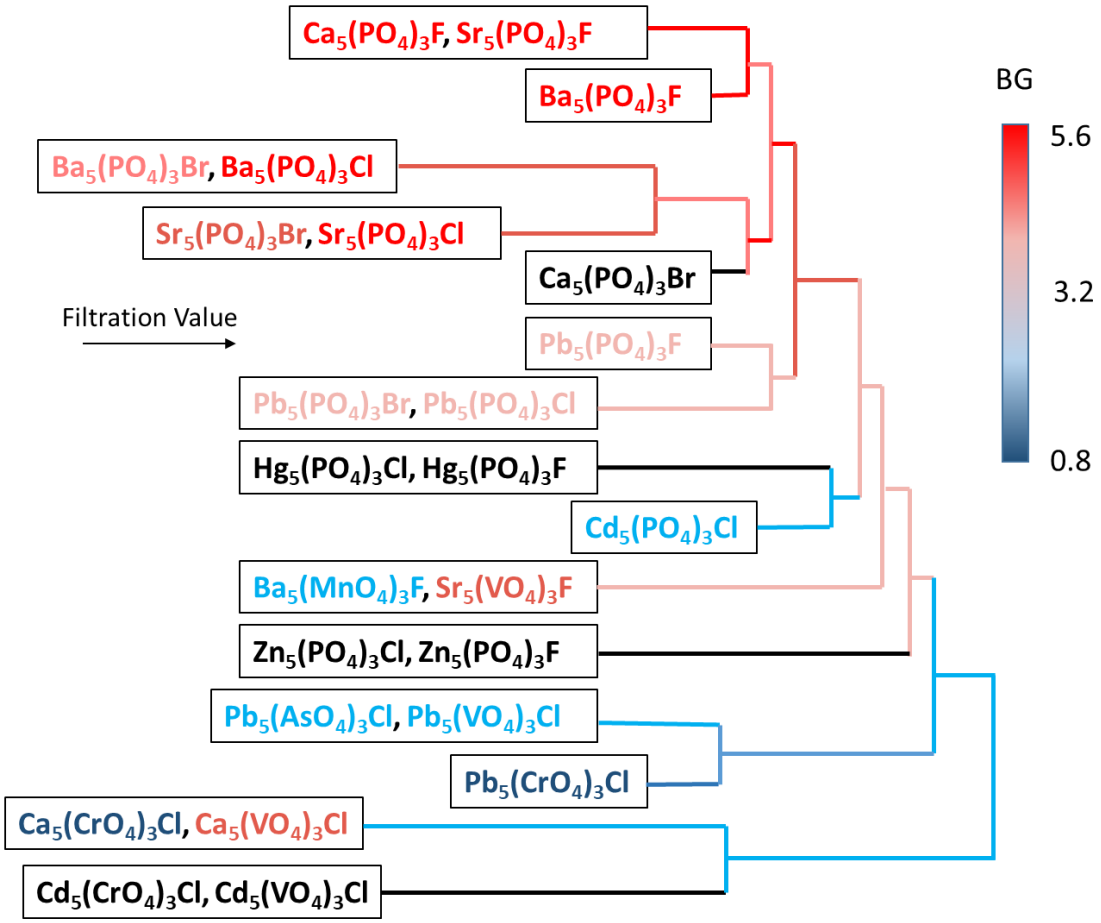

Fig. 3 Dendrogram showing the relationship between bandgap (BG) and the TDA classifications. The compounds in black do not have bandgap values available. With each additional merging of the barcodes as filtration value increases, the collective bandgap decreases. This shows that TDA captures the bandgap without explicit definition. Further, this relationship has been done in an unsupervised manner.

The dendrogram shows a clear trend with bandgap, as the merging of compounds with increasing filtration value results in a decrease in bandgap. The dendrogram was formed following the approach laid out in Figure 2. We see the initial vertical line corresponding with $Ca_5(PO_4)_3F$ and $Sr_5(PO_4)_3F$ having the darkest red line indicating the highest

bandgap. As more compounds merge, the ensuing vertical lines continue to change color to represent lowering bandgap. That is, first a lighter shade of red occurs, and then a blue color. The trends shown here between the barcode evolution and the bandgaps of the corresponding chemistries are anticipated to hold across other computational approaches or even experiments. The estimated behavior of bandgaps is qualitative and therefore subtle differences between calculation methods are all anticipated to be applicable. Additionally, a benefit of this unsupervised learning approach is that any properties can be overlaid to explore a chemistry-property relationship. The relationships will only occur if the correlations are supported by the input data, but in the case of bandgap, the relationship shows up and is strong enough to support the suggested regime of bandgaps for chemistries where data is not available. The analysis was not biased towards bandgap, but rather the analysis was unsupervised. Bandgap was not input into the analysis or explicitly defined, but rather bandgap was compared with the final TDA result. In this way, other properties could be compared to understand what is controlling properties.

As with any data based analysis, the assessment of the impact of input data selection is critical. For this reason, we have repeated the analysis with a difference in descriptor set. In the prior work, we used descriptors spanning both size related descriptors and crystallographic / geometric descriptors. That data required the normalization of the data as the descriptors spanned various units. However, to avoid normalization of data, the analysis was also performed using only size related descriptors (ie. those descriptors with unit of length). In general, the conclusions drawn from the analysis are consistent with the findings from the larger dataset. However, the differences that are present are due to the differences in distortion angles (as the distortion corresponding with building units are not represented in the length only input set). The most noticeable difference is the result of (Ca,Cd)-(Cr,V)-Cl-O. These compounds show up as the last to merge with other compounds when including crystallographic data, but have normal behavior when considering only size related descriptors. Therefore, we can conclude that these four compounds have the largest impact due to distortion between polyhedra of any compounds. Interestingly, $Ca_5(VO_4)_3Cl$ is the most obvious outlier in Figure 3 as it has a bandgap of 3.8 eV which is higher than the other closely merged compounds (for example, $Ca_5(CrO_4)_3Cl$ has bandgap of 0.986 eV). The other compound which does not follow the trends seen is $Sr_5(VO_4)_3F$. Through the application of TDA we are able to identify the compounds which have properties not following the trend with other crystallographic distortions.

To fully understand what is captured in the barcode, we can combine this result linking crystallography and chemistry with property with our prior work which showed the relationship between chemistry and the coordination polyhedra which make up the crystal [14]. As a summary of that work, if we start with an $A_5(BO_4)_3X$ chemistry, the barcode tracks the impact of the bonds on the material behavior. The first change (ie. at the the lowest filtration value) in the groupings captured by the bars is for the A-X bond, which forms the polyhedra of CN = 7. In the provided barcode, the first grouping captures compounds with A = Pb or Sr and X = Br or Cl. Therefore, the barcode is capturing changes in these bonds by representing similarity in four different A-X bond chemistries. We have focused on a subset of initial data for clarity sake, but the grouping at lower filtration values based on A-X bonding is consistent in all different permutations. Initially, no other changes in the material are impacting the classification, beyond the CN = 7 polyhedra. As the barcode grows, the changes in the material are due to the A-O bond. This is where we are classifying additional compounds which modify the A site while other site occupancies remain constant. The consideration of solely A-O bonds is confined to CN = 9 site occupancies. This therefore defines the next building unit required to design an apatite crystal structure, while also serving as the most persistent feature in the crystal. The final change is due to B-O bonding (seen through multiple site occupancies of the B site, including P, Cr, and V). The B-O bond is prevalent in the CN = 4 polyhedra. Therefore, when the crystal structure is developed, the site occupancy of the B-site impacts the CN = 4 polyhedra, which are next in constructing the final crystal.

Based on combining these various pieces of information, we can define which coordination polyhedra appear to be most critical for modifying the bandgap (Figure 4). The CN = 7 polyhedra is the primary factor in compounds with high bandgap. That is, to increase bandgap, modifying the A-X bond is the key to that property engineering. The CN = 4 polyhedra corresponds with the lowest bandgap apatites, and therefore at small bandgap the distortions associated with the B-site dominate the material has lower bandgap.

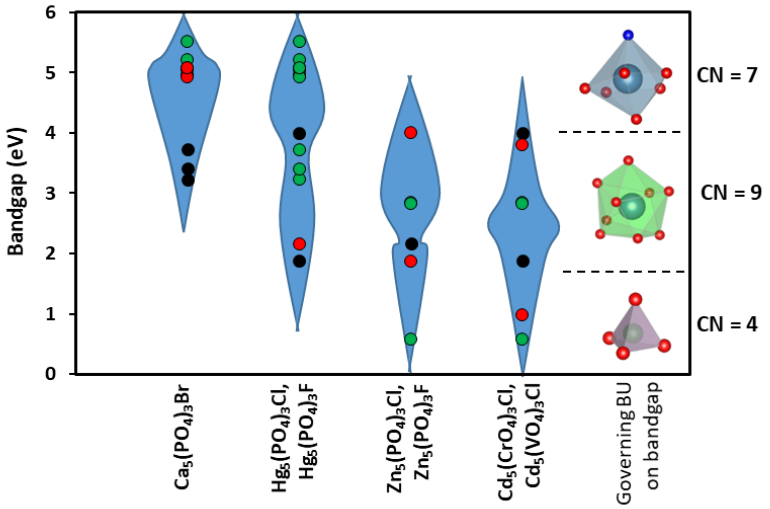

Fig. 4 Estimated range of bandgaps for the compounds without known bandgap values. The range in the violin plot shows the probability of the bandgap occurring, with a broader range representing higher probability. The red points represent the closest connections in the dendrogram of Figure 3, the green points represent next closest connection, and the black points are third closest connection. In our prior work we linked the evolution of TDA with different coordination polyhedra / building units (BUs) governing the connections. Combining that prior work with the bandgap assessment, we are able to define which polyhedra are the primary factor in each bandgap range, with the coordination 7 polyhedra the most critical to have a high bandgap apatite.

In our selection of compounds, there are seven compounds for which we do not have bandgap available. These are shown as black text in Figure 3. By studying the trends in bandgap across the dendrogram, we are able to propose the potential bandgap of these compounds. For comparison, compounds which were not available in Materials Project but were available in AFLOW were used to test this logic, and the result showed good agreement between the bandgap expected based on the merging of compounds with their filtration values. Figure 4 shows the anticipated bandgaps for these compounds. The data has been presented through a violin plot, which acknowledges the uncertainty in this classification. That is, we are not defining explicit models or following a proper regression analysis, and this is done through a qualitative comparison. The red dots are the compounds which are most closely connected with the compounds, the green dots are the next connections and the black dots are third level connections. For example, in the case of $Ca_5(PO_4)_3Br$, the compounds it first merges with are $Ba_5(PO_4)_3Br$, $Ba_5(PO_4)_3Cl$, $Sr_5(PO_4)_3Br$, and $Sr_5(PO_4)_3Cl$. The bandgaps for those compounds are shown as the red dots for $Ca_5(PO_4)_3Br$. In defining the likely bandgap, the red dots have a higher weighting in defining the probability. The violin plot having the widest region near those points, while the range is narrow near the black dots. The two Hg compounds are on the same bar, and therefore we are unable to differentiate them in terms of expected bandgap. For this reason, the seven compounds are represented by four figures.

This result provides an estimate of likely bandgap, with uncertainty accounted for. The compounds are shown in the order of anticipated bandgaps, going from highest to lowest. This demonstrates the use of TDA for not only representing the data, and in particular the connectivity of data, but also for design implications. The data input into the analysis is largely based on size effects and therefore a large number of compounds could be added to this as the

descriptor requirements are minimal. This work is the first time of using TDA as a stand-alone design tool. That is, TDA has been used to develop additional parameters to improve regression models [24-26] but has not been used alone as a design tool. The reason why this is important is because the analysis is done in an unsupervised manner and therefore this approach can be applied even with relatively small data as we are not biasing the analysis.

While we have shown the broad applicability of this TDA approach and the benefits coming from it, there are some restrictions. As with all data driven approaches, the results and conclusions are dependent on the input data. This raises an additional challenge in TDA as we do not have an exact ranking of the impact of the descriptors on the model, beyond running the analysis with many different permutations of the input data. This supports the integration of TDA with other approaches that provide more descriptor assessment, such as principal component analysis. This can also be used to select the input to minimize the redundancy of the descriptors. An additional important consideration for addressing this limitation is to assess the results in terms of physical significance. In this paper, we have done that through the analysis of bandgap, finding that the analysis is capturing physically meaningful trends. An additional limitation in the TDA approach is that the trends are largely qualitative; however, our interest here is in guiding future material design and experiments, and therefore a categorical prediction provides sufficient screening of the design space.

As noted by Thompson and Down [27], crystal chemistry has traditionally considered the analysis of anion-cation interactions as critical to the understanding of stability and a crystal structure's response to temperature, pressure, and composition. A pure hard-sphere model of crystal geometry is not sufficient in addressing the link between the chemical identity of an atom and the geometry (coordination) of its environment, as described by bond lengths and bond. Recently there have been reports that have reinforced the strong link between coordination polyhedra and bond distortions and band gap in complex inorganic structures as suggested by our TDA findings. For example, ten Kate et. al. [28] through a detailed empirical study of experimental data, found that variations in Si/N ratio influenced the coordination number of N by changing its effective charge; which in turn influenced the bandgap.

Conclusions

In this paper, we have introduced the value of using topological data analysis as an inference tool to identify local chemical environments that can serve as "digital landmarks" to uncover new and unexplored chemistry-property associations. By linking stoichiometry to bond specific characteristics in a chemically complex unit cell coupled to a framework that permits one to track the dynamics of influence of descriptors on properties, we can now rapidly and autonomously survey connectivity in data. The harnessing of the geometry/ topology of high dimensional data with the statistical characteristics of data provides a powerful foundation for a 'data' guided discovery platform.

Acknowledgements

We gratefully acknowledge support from National Science Foundation (NSF) DIBBs program, award number 1640867. KR also acknowledges the Erich Bloch Endowment at the University at Buffalo- State University of New York.

Conflict of Interest

The authors do not have any conflict of interests to report.

Data Availability

The data and codes used in this study are available through the 'Materials Data Engineering' web portal (www.madeatub.buffalo.edu/MaDE@UB).

References

[1] R. Berger, S. Lee, J. Johnson, B. Nebgen, F. Sha, J. Xu, Eur. J. Chem. 14, 3908 (2008)

[2] R. Berger, S. Lee, J. Johnson, B. Nebgen, A.C.-Y. So, Eur. J. Chem., 14, 6617 (2008)
[3] A.L. Mackay, Comp. Maths. Appl. B 12, 21 (1966)
[4] A.L. Mackay, Acta Cryst. A 30, 440 (1974)
[5] A.L. Mackay, Acta Cryst. A 33, 212 (1977)
[6] A.L. Mackay, Proc. of the 1st Int. Symposium on Form, eds. Y. Kato, R. Takaki, J. Toriwaki, 615 (1986)
[7] J.H.E. Cartwright, A.L. Mackay, Phil. Trans. R. Soc. A 370, 2807 (2012)
[8] A.L. Mackay, J. Kinowski, Comp. Maths. Appls. 21B, 803 (1986)
[9] V.A. Blatov, G.D. Ilyushin, D.M. Proserpio, Inorg. Chem. 49, 1811 (2010)
[10] F.C. Hawthorne, Phys. Chem. Minerals 39, 841 (2012)
[11] H.P. Beck, Z. Kristallogr. 229, 473 (2014)
[12] M. Offroy, L. Duponchel, Analytica Chim. Acta, 910, 1 (2016)
[13] G. Carlsson, Curr. Opinion Syst. Bio. 1, 109 (2017)
[14] S. Broderick, R. Dongol, T. Zhang, K. Rajan, Sci. Repts. 11, 11599 (2021)
[15] S. Bhattacharya, R. Ghrist, V. Kumar, IEEE Trans. Robotics 31, 578 (2015)
[16] G. Carlsson, T. Ishkhanov, V. d. Silva, A. Zomorodian, Intl. J. Comp. Vision 76, 1 (2008)
[17] I. Donato, M. Gori, M. Pettini, G. Petri, S. De Nigris, R. Franzosi, F. Vaccarino, Phys. Rev. E 93, 052138 (2016)
[18] H. Edelsbrunner, D. Letscher, A. Zomorodian, Discrete Comp. Geom. 28, 511 (2002)
[19] J. Townsend, C.P. Micucci, J.H. Hymel, V. Maroulas, K.D. Vogiatzis, Nat. Comm. 11, 3230 (2020)
[20] P.V. Balachandran, K. Rajan, Acta Cryst. B68, 24 (2012)
[21] C.S. Kong, W. Luo, S. Arapan, P. Villars, S. Iwata, R. Ahuja, K. Rajan, Chem. Inf. Modeling 52, 1812 (2012)
[22] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, APL Matls 1, 011002 (2013)
[23] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, Comp. Mat. Sci. 58, 218 (2012)
[24] A.S. Krishnapriyan, M. Haranczyk, D. Morozv, J. Phys. Chem. C 124, 9360 (2020)
[25] X. Chen, D. Chen, M. Weng, Y. Jiang, G-W. Wei, F. Pan, J. Phys. Chem. Lett. 11, 4392 (2020)
[26] Y. Jiang, D. Chen, X. Chen, T. Li, G-W. Wei, F. Pan, npj Comp. Matls. 28, 1 (2021)
[27] R.M. Thompson, R.T. Downs, Acta Cryst. B 57, 119 (2001)
[28] O.M. ten Kate, Z. Zhang, H. T.Hintzen, J. Mater. Chem. C 5, 11504 (2017)